Fundamentals of
Engineering Electromagnetics

© 2006 by Taylor & Francis Group, LLC

VISIT…



Fundamentals of
Engineering Electromagnetics

edited by
Rajeev Bansal



Taylor & Francis
Taylor & Francis Group
Boca Raton London New York

A CRC title, part of the Taylor & Francis imprint, a member of the Taylor & Francis Group, the academic division of T&F Informa plc.

© 2006 by Taylor & Francis Group, LLC

The material was previously published in *The Handbook of Engineering Electromagnetics* © Taylor & Francis 2004.

Published in 2006 by
CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2006 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-8493-7360-3 (Hardcover)
International Standard Book Number-13: 978-0-8493-7360-2 (Hardcover)
Library of Congress Card Number 2005058201

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Bansal, Rajeev.
Fundamentals of engineering electromagnetics / Rajeev Bansal.
p. cm.
Includes bibliographical references and index.
ISBN 0-8493-7360-3 (alk. paper)
1. Electromagnetism. I. Title.

QC760.B2653 2006
621.3--dc22 2005058201

informa
Taylor & Francis Group
is the Academic Division of Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

© 2006 by Taylor & Francis Group, LLC

To the memory of my parents

© 2006 by Taylor & Francis Group, LLC

Preface

Aim

This volume, derived from the *Handbook of Engineering Electromagnetics* (2004), is intended as a desk reference for the fundamentals of engineering electromagnetics. Because electromagnetics provides the underpinnings for many technological fields such as wireless communications, fiber optics, microwave engineering, radar, electromagnetic compatibility, material science, and biomedicine, there is a great deal of interest and need for training in the concepts of engineering electromagnetics. Practicing engineers in these diverse fields must understand how electromagnetic principles can be applied to the formulation and solution of actual engineering problems.

Fundamentals of Engineering Electromagnetics should serve as a bridge between standard textbooks in electromagnetic theory and specialized references such as handbooks on radar or wireless communications. While textbooks are comprehensive in terms of the theoretical development of the subject matter, they are usually deficient in the application of that theory to practical applications. Specialized handbooks, on the other hand, often provide detailed lists of formulas, tables, and graphs, but do not offer the insight needed to appreciate the underlying physical concepts. This volume will permit a practicing engineer/scientist to:

Review the necessary electromagnetic theory.
Gain an appreciation for the key electromagnetic terms and parameters.
Learn how to apply the theory to formulate engineering problems.
Obtain guidance to the specialized literature for additional details.

Scope

Because *Fundamentals of Engineering Electromagnetics* is intended to be useful to engineers engaged in electromagnetic applications in a variety of professional settings, the coverage of topics is correspondingly broad, including Maxwell equations, static fields, electromagnetic induction, waves, transmission lines, waveguides, antennas, and

© 2006 by Taylor & Francis Group, LLC

electromagnetic compatibility. Pertinent *data* in the form of tables and graphs has been provided within the context of the subject matter. In addition, Appendices A and B are brief compilations of important electromagnetic constants and units, respectively. Finally, Appendix C is intended as a convenient tutorial on vector analysis and coordinate systems.

© 2006 by Taylor & Francis Group, LLC

Acknowledgments

First and foremost, I would like to thank all the contributors, whose hard work is reflected in the pages of this volume. My editors at Taylor & Francis, specially Mr. Taisuke Soda, have provided valuable help and advice throughout the project. I would like to thank Mr. Anthony Palladino for his help in preparing the manuscript of Appendix C. Finally, I would like to express my gratitude to my family for their unfailing support and encouragement.

© 2006 by Taylor & Francis Group, LLC

Editor

Rajeev Bansal received his Ph.D. in Applied Physics from Harvard University in 1981. Since then he has taught and conducted research in the area of applied electromagnetics at the University of Connecticut, where he is currently a professor of electrical engineering. His technical contributions include the edited volume *Handbook of Engineering Electromagnetics* (2004), two coauthored book chapters on submarine antennas (2005) and semiconductor dipole antennas (1986), two patents (1989 and 1993), and over 75 journal/conference papers. Dr. Bansal has served on the editorial boards of *Int. J. of RF and Microwave Computer-Aided Engineering*, *Journal of Electromagnetic Waves and Applications*, *Radio Science*, *IEEE Antennas and Propagation Magazine*, and *IEEE Microwave Magazine*. He is a member of the Electromagnetics Academy and the Technical Coordinating Committee of the IEEE Microwave Theory & Techniques Society. He has served as a consultant to the Naval Undersea Warfare Center, Newport, RI.

© 2006 by Taylor & Francis Group, LLC

Contributors

Christo Christopoulos *University of Nottingham, Nottingham, England*

Kenneth R. Demarest *The University of Kansas, Lawrence, Kansas*

Mark N. Horenstein *Boston University, Boston, Massachusetts*

David R. Jackson *University of Houston, Houston, Texas*

Mohammad Kolbehdari *Intel Corporation, Hillsboro, Oregon*

Branko D. Popović[†] *University of Belgrade, Belgrade, Yugoslavia*

Milica Popović *McGill University, Montreal, Quebec, Canada*

Zoya Popović *University of Colorado, Boulder, Colorado*

N. Narayana Rao *University of Illinois at Urbana-Champaign, Urbana, Illinois*

Matthew N. O. Sadiku *Prairie View A&M University, Prairie View, Texas*

David Thiel *Griffith University, Nathan, Queensland, Australia*

Andreas Weisshaar *Oregon State University, Corvallis, Oregon*

Jeffrey T. Williams *University of Houston, Houston, Texas*

Donald R. Wilton *University of Houston, Houston, Texas*

[†] *Deceased.*

© 2006 by Taylor & Francis Group, LLC

Contents

1. Fundamentals of Engineering Electromagnetics Revisited 1
N. Narayana Rao

2. Applied Electrostatics 53
Mark N. Horenstein

3. Magnetostatics 89
Milica Popović, Branko D. Popović†, and Zoya Popović

4. Electromagnetic Induction 123
Milica Popović, Branko D. Popović†, and Zoya Popović

5. Wave Propagation 163
Mohammad Kolbehdari and Matthew N. O. Sadiku

6. Transmission Lines 185
Andreas Weisshaar

7. Waveguides and Resonators 227
Kenneth R. Demarest

8. Antennas: Fundamentals 255
David Thiel

9. Antennas: Representative Types 277
David R. Jackson, Jeffery T. Williams, and Donald R. Wilton

10. Electromagnetic Compatibility 347
Christos Christopoulos

Appendix A: Some Useful Constants 377
Appendix B: Some Units and Conversions 379
Appendix C: Review of Vector Analysis and Coordinate Systems 381

© 2006 by Taylor & Francis Group, LLC

1

Fundamentals of Engineering Electromagnetics Revisited

N. Narayana Rao
University of Illinois at Urbana-Champaign
Urbana, Illinois

In this chapter, we present in a nutshell the fundamental aspects of engineering electromagnetics from the view of looking back in a reflective fashion at what has already been learned in undergraduate electromagnetics courses as a novice. The first question that comes to mind in this context is what constitutes the fundamentals of engineering electromagnetics. If the question is posed to several individuals, it is certain that they will come up with sets of topics, not necessarily the same or in the same order, but all containing the topic of Maxwell's equations at some point in the list, ranging from the beginning to the end of the list. In most cases, the response is bound to depend on the manner in which the individual was first exposed to the subject. Judging from the contents of the vast collection of undergraduate textbooks on electromagnetics, there is definitely a heavy tilt toward the traditional, or historical, approach of beginning with statics and culminating in Maxwell's equations, with perhaps an introduction to waves. Primarily to provide a more rewarding understanding and appreciation of the subject matter, and secondarily owing to my own fascination resulting from my own experience as a student, a teacher, and an author [1–7] over a few decades, I have employed in this chapter the approach of beginning with Maxwell's equations and treating the different categories of fields as solutions to Maxwell's equations. In doing so, instead of presenting the topics in an unconnected manner, I have used the thread of statics–quasistatics–waves to cover the fundamentals and bring out the frequency behavior of physical structures at the same time.

1.1. FIELD CONCEPTS AND CONSTITUTIVE RELATIONS

1.1.1. Lorentz Force Equation

A region is said to be characterized by an electric field if a particle of charge q moving with a velocity $\mathbf{v}$ experiences a force $\mathbf{F}_e$, independent of $\mathbf{v}$. The force, $\mathbf{F}_e$, is given by

$$\mathbf{F}_e = q\mathbf{E} \tag{1.1}$$

© 2006 by Taylor & Francis Group, LLC

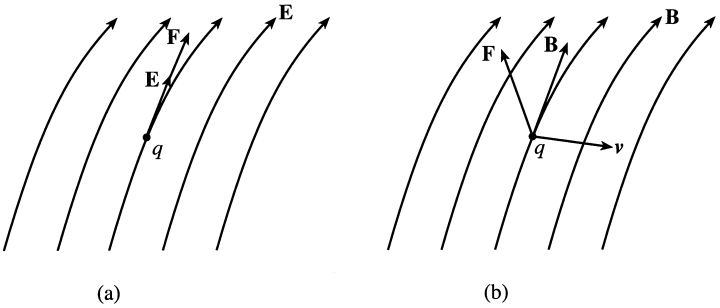


Figure 1.1 Illustrates that (a) the electric force is parallel to **E** but (b) the magnetic force is perpendicular to **B**.

where **E** is the electric field intensity, as shown in Fig. 1.1a. We note that the units of **E** are newtons per coulomb (N/C). Alternate and more commonly used units are volts per meter (V/m), where a volt is a newton-meter per coulomb. The line integral of **E** between two points A and B in an electric field region, $\int_A^B \mathbf{E}\cdot d\mathbf{l}$, has the meaning of voltage between A and B. It is the work per unit charge done by the field in the movement of the charge from A to B. The line integral of **E** around a closed path C is also known as the *electromotive force* (emf) around C.

If the charged particle experiences a force which depends on **v**, then the region is said to be characterized by a magnetic field. The force, $\mathbf{F}_m$, is given by

$$\mathbf{F}_m = q\mathbf{v} \times \mathbf{B} \tag{1.2}$$

where **B** is the magnetic flux density. We note that the units of **B** are newtons/(coulomb-meter per second), or (newton-meter per coulomb) × (seconds per square meter), or volt-seconds per square meter. Alternate and more commonly used units are webers per square meter (Wb/m^2) or tesla (T), where a weber is a volt-second. The surface integral of **B** over a surface S, $\int_S \mathbf{B}\cdot d\mathbf{S}$, is the magnetic flux (Wb) crossing the surface.

Equation (1.2) tells us that the magnetic force is proportional to the magnitude of **v** and orthogonal to both **v** and **B** in the right-hand sense, as shown in Fig. 1.1b. The magnitude of the force is $qvB \sin\alpha$, where α is the angle between **v** and **B**. Since the force is normal to **v**, there is no acceleration along the direction of motion. Thus the magnetic field changes only the direction of motion of the charge and does not alter the kinetic energy associated with it.

Since current flow in a wire results from motion of charges in the wire, a wire of current placed in a magnetic field experiences a magnetic force. For a differential length $d\mathbf{l}$ of a wire of current I placed in a magnetic field **B**, this force is given by

$$d\mathbf{F}_m = I\, d\mathbf{l} \times \mathbf{B} \tag{1.3}$$

as shown in Fig. 1.2.

Combining Eqs. (1.1) and (1.2), we obtain the expression for the total force $\mathbf{F} = \mathbf{F}_e + \mathbf{F}_m$, experienced by a particle of charge q moving with a velocity **v** in a region of

© 2006 by Taylor & Francis Group, LLC

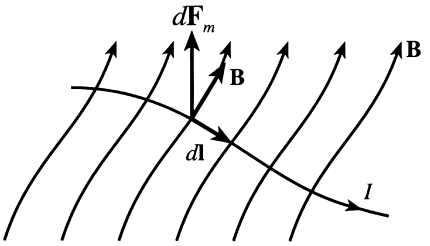


Figure 1.2 Force experienced by a current element in a magnetic field.

electric and magnetic fields, **E** and **B**, respectively, as

$$\begin{aligned}\mathbf{F} &= q\mathbf{E} + q\mathbf{v} \times \mathbf{B} \\ &= q(\mathbf{E} + \mathbf{v} \times \mathbf{B})\end{aligned} \tag{1.4}$$

Equation (1.4) is known as the *Lorentz force equation*.

1.1.2. Material Parameters and Constitutive Relations

The vectors **E** and **B** are the fundamental field vectors that define the force acting on a charge moving in an electromagnetic field, as given by the Lorentz force Eq. (1.4). Two associated field vectors **D** and **H**, known as the *electric flux density* (or the *displacement flux density*) and the *magnetic field intensity*, respectively, take into account the dielectric and magnetic properties, respectively, of material media. Materials contain charged particles that under the application of external fields respond giving rise to three basic phenomena known as *conduction*, *polarization*, and *magnetization*. Although a material may exhibit all three properties, it is classified as a *conductor*, a *dielectric*, or a *magnetic* material depending upon whether conduction, polarization, or magnetization is the predominant phenomenon. While these phenomena occur on the atomic or "microscopic" scale, it is sufficient for our purpose to characterize the material based on "macroscopic" scale observations, that is, observations averaged over volumes large compared with atomic dimensions.

In the case of conductors, the effect of conduction is to produce a current in the material known as the *conduction current*. Conduction is the phenomenon whereby the free electrons inside the material move under the influence of the externally applied electric field with an average velocity proportional in magnitude to the applied electric field, instead of accelerating, due to the frictional mechanism provided by collisions with the atomic lattice. For linear isotropic conductors, the conduction current density, having the units of amperes per square meter (A/m^2), is related to the electric field intensity in the manner

$$\mathbf{J}_c = \sigma\mathbf{E} \tag{1.5}$$

where σ is the conductivity of the material, having the units siemens per meter (S/m). In semiconductors, the conductivity is governed by not only electrons but also holes.

While the effect of conduction is taken into account explicitly in the electromagnetic field equations through Eq. (1.5), the effect of polarization is taken into account implicitly

© 2006 by Taylor & Francis Group, LLC

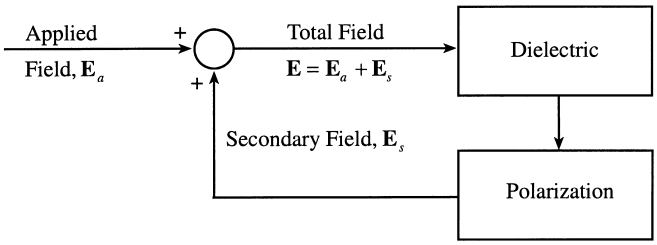


Figure 1.3 Illustrates the effect of polarization in a dielectric material.

through the relationship between **D** and **E**, which is given by

$$\mathbf{D} = \varepsilon \mathbf{E} \tag{1.6}$$

for linear isotropic dielectrics, where ε is the *permittivity* of the material having the units coulomb squared per newton-squared meter, commonly known as *farads per meter* (F/m), where a *farad* is a coulomb square per newton-meter.

Polarization is the phenomenon of creation and net alignment of electric dipoles, formed by the displacements of the centroids of the electron clouds of the nuclei of the atoms within the material, along the direction of an applied electric field. The effect of polarization is to produce a secondary field that acts in superposition with the applied field to cause the polarization. Thus the situation is as depicted in Fig. 1.3. To implicitly take this into account, leading to Eq. (1.6), we begin with

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \tag{1.7}$$

where ε_0 is the permittivity of free space, having the numerical value 8.854×10^{-12}, or approximately $10^{-9}/36\pi$, and **P** is the polarization vector, or the dipole moment per unit volume, having the units (coulomb-meters) per cubic meter or coulombs per square meter. Note that this gives the units of coulombs per square meter for **D**. The term $\varepsilon_0 \mathbf{E}$ accounts for the relationship between **D** and **E** if the medium were free space, and the quantity **P** represents the effect of polarization. For linear isotropic dielectrics, **P** is proportional to **E** in the manner

$$\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E} \tag{1.8}$$

where χ_e, a dimensionless quantity, is the electric susceptibility, a parameter that signifies the ability of the material to get polarized. Combining Eqs. (1.7) and (1.8), we have

$$\begin{aligned} \mathbf{D} &= \varepsilon_0(1 + \chi_e)\mathbf{E} \\ &= \varepsilon_0 \varepsilon_r \mathbf{E} \\ &= \varepsilon \mathbf{E} \end{aligned} \tag{1.9}$$

where ε_r $(= 1 + \chi_e)$ is the relative permittivity of the material.

© 2006 by Taylor & Francis Group, LLC

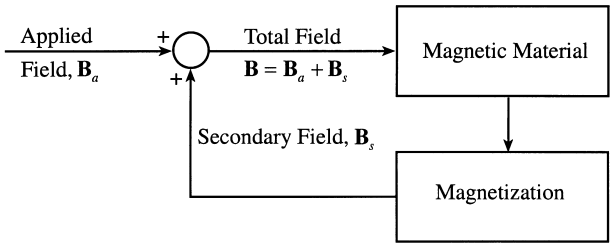


Figure 1.4 Illustrates the effect of magnetization in a magnetic material.

In a similar manner, the effect of magnetization is taken into account implicitly through the relationship between **H** and **B**, which is given by

$$\mathbf{H} = \frac{\mathbf{B}}{\mu} \tag{1.10}$$

for linear isotropic magnetic materials, where μ is the permeability of the material, having the units newtons per ampere squared, commonly known as *henrys per meter* (H/m), where a *henry* is a newton-meter per ampere squared.

Magnetization is the phenomenon of net alignment of the axes of the magnetic dipoles, formed by the electron orbital and spin motion around the nuclei of the atoms in the material, along the direction of the applied magnetic field. The effect of magnetization is to produce a secondary field that acts in superposition with the applied field to cause the magnetization. Thus the situation is as depicted in Fig. 1.4. To implicitly take this into account, we begin with

$$\mathbf{B} = \mu_0\mathbf{H} + \mu_0\mathbf{M} \tag{1.11}$$

where μ_0 is the permeability of free space, having the numerical value $4\pi \times 10^{-7}$, and **M** is the magnetization vector or the magnetic dipole moment per unit volume, having the units (ampere-square meters) per cubic meter or amperes per meter. Note that this gives the units of amperes per square meter for **H**. The term $\mu_0\mathbf{H}$ accounts for the relationship between **H** and **B** if the medium were free space, and the quantity $\mu_0\mathbf{M}$ represents the effect of magnetization. For linear isotropic magnetic materials, **M** is proportional to **H** in the manner

$$\mathbf{M} = \chi_m\mathbf{H} \tag{1.12}$$

where χ_m, a dimensionless quantity, is the magnetic susceptibility, a parameter that signifies the ability of the material to get magnetized. Combining Eqs. (1.11) and 1.12),

© 2006 by Taylor & Francis Group, LLC

we have

$$\begin{aligned}\mathbf{H} &= \frac{\mathbf{B}}{\mu_0(1+\chi_m)} \\ &= \frac{\mathbf{B}}{\mu_0\mu_r} \\ &= \frac{\mathbf{B}}{\mu}\end{aligned} \tag{1.13}$$

where $\mu_r\ (=1+\chi_m)$ is the relative permeability of the material.

Equations (1.5), (1.6), and (1.10) are familiarly known as the *constitutive relations*, where σ, ε, and μ are the material parameters. The parameter σ takes into account explicitly the phenomenon of conduction, whereas the parameters ε and μ take into account implicitly the phenomena of polarization and magnetization, respectively.

The constitutive relations, Eqs. (1.5), (1.6), and (1.10), tell us that $\mathbf{J}_c$ is parallel to $\mathbf{E}$, $\mathbf{D}$ is parallel to $\mathbf{E}$, and $\mathbf{H}$ is parallel to $\mathbf{B}$, independent of the directions of the field vectors. For anisotropic materials, the behavior depends upon the directions of the field vectors. The constitutive relations have then to be written in matrix form. For example, in an anisotropic dielectric, each component of $\mathbf{P}$ and hence of $\mathbf{D}$ is in general dependent upon each component of $\mathbf{E}$. Thus, in terms of components in the Cartesian coordinate system, the constitutive relation is given by

$$\begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \tag{1.14}$$

or, simply by

$$[\mathbf{D}] = [\varepsilon][\mathbf{E}] \tag{1.15}$$

where $[\mathbf{D}]$ and $[\mathbf{E}]$ are the column matrices consisting of the components of $\mathbf{D}$ and $\mathbf{E}$, respectively, and $[\varepsilon]$ is the permittivity matrix (tensor of rank 2) containing the elements ε_{ij}, $i=1,\ 2,\ 3$ and $j=1,\ 2,\ 3$. Similar relationships hold for anisotropic conductors and anisotropic magnetic materials.

Since the permittivity matrix is symmetric, that is, $\varepsilon_{ij}=\varepsilon_{ji}$, from considerations of energy conservation, an appropriate choice of the coordinate system can be made such that some or all of the nondiagonal elements are zero. For a particular choice, all of the nondiagonal elements can be made zero so that

$$[\varepsilon] = \begin{bmatrix} \varepsilon_1 & 0 & 0 \\ 0 & \varepsilon_2 & 0 \\ 0 & 0 & \varepsilon_3 \end{bmatrix} \tag{1.16}$$

Then

$$D_{x'} = \varepsilon_1 E_{x'} \tag{1.17a}$$

$$D_{y'} = \varepsilon_2 E_{y'} \tag{1.17b}$$

© 2006 by Taylor & Francis Group, LLC

$$D_{z'} = \varepsilon_3 E_{z'} \tag{1.17c}$$

so that **D** and **E** are parallel when they are directed along the coordinate axes, although with different values of *effective permittivity*, that is, ratio of **D** to **E**, for each such direction. The axes of the coordinate system are then said to be the *principal axes* of the medium. Thus when the field is directed along a principal axis, the anisotropic medium can be treated as an isotropic medium of permittivity equal to the corresponding effective permittivity.

1.2. MAXWELL'S EQUATIONS, BOUNDARY CONDITIONS, POTENTIALS, AND POWER AND ENERGY

1.2.1. Maxwell's Equations in Integral Form and the Law of Conservation of Charge

In Sec. 1.1, we introduced the different field vectors and associated constitutive relations for material media. The electric and magnetic fields are governed by a set of four laws, known as *Maxwell's equations*, resulting from several experimental findings and a purely mathematical contribution. Together with the constitutive relations, Maxwell's equations form the basis for the entire electromagnetic field theory. In this section, we shall consider the time variations of the fields to be arbitrary and introduce these equations and an auxiliary equation in the time domain form. In view of their experimental origin, the fundamental form of Maxwell's equations is the integral form. In the following, we shall first present all four Maxwell's equations in integral form and the auxiliary equation, the law of conservation of charge, and then discuss several points of interest pertinent to them. It is understood that all field quantities are real functions of position and time; that is, $\mathbf{E} = \mathbf{E}(\mathbf{r}, t) = \mathbf{E}(x, y, z, t)$, etc.

Faraday's Law

Faraday's law is a consequence of the experimental finding by Michael Faraday in 1831 that a time-varying magnetic field gives rise to an electric field. Specifically, the electromotive force around a closed path C is equal to the negative of the time rate of increase of the magnetic flux enclosed by that path, that is,

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{S} \tag{1.18}$$

where S is any surface bounded by C, as shown, for example, in Fig. 1.5.

Ampere's Circuital Law

Ampere's circuital law is a combination of an experimental finding of Oersted that electric currents generate magnetic fields and a mathematical contribution of Maxwell that time-varying electric fields give rise to magnetic fields. Specifically, the magnetomotive force (mmf) around a closed path C is equal to the sum of the current enclosed by that path due

© 2006 by Taylor & Francis Group, LLC

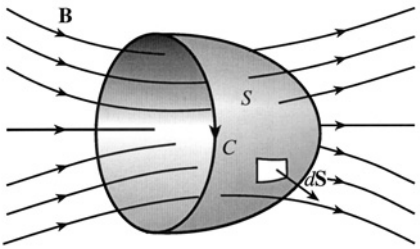


Figure 1.5 Illustrates Faraday's law.

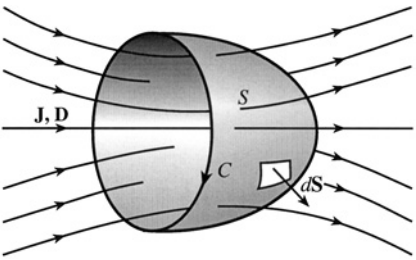


Figure 1.6 Illustrates Ampere's circuital law.

to actual flow of charges and the displacement current due to the time rate of increase of the electric flux (or displacement flux) enclosed by that path; that is,

$$\oint_C \mathbf{H}\cdot d\mathbf{l} = \int_S \mathbf{J}\cdot d\mathbf{S} + \frac{d}{dt}\int_S \mathbf{D}\cdot d\mathbf{S} \tag{1.19}$$

where S is any surface bounded by C, as shown, for example, in Fig. 1.6.

Gauss' Law for the Electric Field

Gauss' law for the electric field states that electric charges give rise to electric field. Specifically, the electric flux emanating from a closed surface S is equal to the charge enclosed by that surface, that is,

$$\oint_S \mathbf{D}\cdot d\mathbf{S} = \int_V \rho\, dv \tag{1.20}$$

where V is the volume bounded by S, as shown, for example, in Fig. 1.7. In Eq. (1.20), the quantity ρ is the volume charge density having the units coulombs per cubic meter (C/m^3).

© 2006 by Taylor & Francis Group, LLC

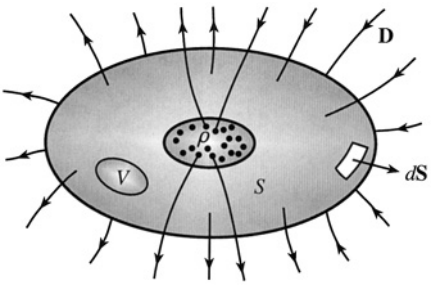


Figure 1.7 Illustrates Gauss' law for the electric field.

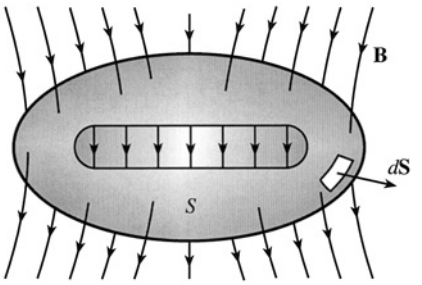


Figure 1.8 Illustrates Gauss' law for the magnetic field.

Gauss' Law for the Magnetic Field

Gauss' law for the magnetic field states that the magnetic flux emanating from a closed surface S is equal to zero, that is,

$$\oint_S \mathbf{B}\cdot d\mathbf{S} = 0 \tag{1.21}$$

Thus, whatever magnetic flux enters (or leaves) a certain part of the closed surface must leave (or enter) through the remainder of the closed surface, as shown, for example, in Fig. 1.8.

Law of Conservation of Charge

An auxiliary equation known as the *law of conservation of charge* states that the current due to flow of charges emanating from a closed surface S is equal to the time rate of decrease of the charge inside the volume V bounded by that surface, that is,

$$\oint_S \mathbf{J}\cdot d\mathbf{S} = -\frac{d}{dt}\int_V \rho\, dv$$

or

$$\oint_S \mathbf{J}\cdot d\mathbf{S} + \frac{d}{dt}\int_V \rho\, dv = 0 \tag{1.22}$$

© 2006 by Taylor & Francis Group, LLC

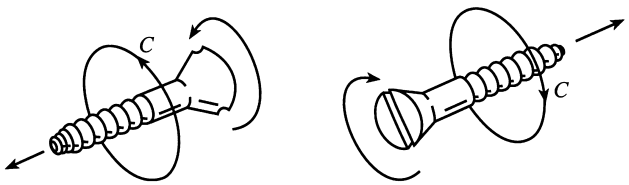


Figure 1.9 Right-hand-screw-rule convention.

There are certain procedures and observations of interest pertinent to Eqs. (1.18)–(1.22), as follows.

1. The direction of the infinitesimal surface vector $d\mathbf{S}$ in Figs. 1.5 and 1.6 denotes that the magnetic flux and the displacement flux, respectively, are to be evaluated in accordance with the right-hand screw rule (RHS rule), that is, in the sense of advance of a right-hand screw as it is turned around C in the sense of C, as shown in Fig. 1.9. The RHS rule is a convention that is applied consistently for all electromagnetic field laws involving integration over surfaces bounded by closed paths.
2. In evaluating the surface integrals in Eqs. (1.18) and (1.19), any surface S bounded by C can be employed. In addition in Eq. (1.19), the same surface S must be employed for both surface integrals. This implies that the time derivative of the magnetic flux through all possible surfaces bounded by C is the same in order for the emf around C to be unique. Likewise, the sum of the current due to flow of charges and the displacement current through all possible surfaces bounded C is the same in order for the mmf around C to be unique.
3. The minus sign on the right side of Eq. (1.18) tells us that when the magnetic flux enclosed by C is increasing with time, the induced voltage is in the sense opposite to that of C. If the path C is imagined to be occupied by a wire, then a current would flow in the wire that produces a magnetic field so as to oppose the increasing flux. Similar considerations apply for the case of the magnetic flux enclosed by C decreasing with time. These are in accordance with Lenz' law, which states that the sense of the induced emf is such that any current it produces tends to oppose the change in the magnetic flux producing it.
4. If loop C contains more than one turn, such as in an N-turn coil, then the surface S bounded by C takes the shape of a spiral ramp, as shown in Fig. 1.10. For a tightly wound coil, this is equivalent to the situation in which N separate, identical, single-turn loops are stacked so that the emf induced in the N-turn coil is N times the emf induced in one turn. Thus, for an N-turn coil,

$$\text{emf} = -N\frac{d\psi}{dt} \tag{1.23}$$

   where ψ is the magnetic flux computed as though the coil is a one-turn coil.
5. Since magnetic force acts perpendicular to the motion of a charge, the magnetomotive (mmf) force, that is, $\oint_C \mathbf{H}\cdot d\mathbf{l}$, does not have a physical meaning similar to that of the electromotive force. The terminology arises purely from analogy with electromotive force for $\oint_C \mathbf{E}\cdot d\mathbf{l}$.

© 2006 by Taylor & Francis Group, LLC

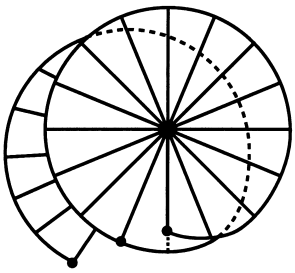


Figure 1.10 Two-turn loop.

6. The charge density ρ in Eq. (1.20) and the current density **J** in Eq. (1.19) pertain to true charges and currents, respectively, due to motion of true charges. They do not pertain to charges and currents resulting from the polarization and magnetization phenomena, since these are implicitly taken into account by the formulation of these two equations in terms of **D** and **H**, instead of in terms of **E** and **B**.
7. The displacement current, $d(\int_S \mathbf{D}\cdot d\mathbf{S})/dt$ is not a true current, that is, it is not a current due to actual flow of charges, such as in the case of the conduction current in wires or a convection current due to motion of a charged cloud in space. Mathematically, it has the units of $d[(\mathrm{C/m^2}) \times \mathrm{m^2}]/dt$ or amperes, the same as the units for a true current, as it should be. Physically, it leads to the same phenomenon as a true current does, even in free space for which **P** is zero, and **D** is simply equal to $\varepsilon_0\mathbf{E}$. Without it, the uniqueness of the mmf around a given closed path C is not ensured. In fact, Ampere's circuital law in its original form did not contain the displacement current term, thereby making it valid only for the static field case. It was the mathematical contribution of Maxwell that led to the modification of the original Ampere's circuital law by the inclusion of the displacement current term. Together with Faraday's law, this modification in turn led to the theoretical prediction by Maxwell of the phenomenon of electromagnetic wave propagation in 1864 even before it was confirmed experimentally 23 years later in 1887 by Hertz.
8. The observation concerning the time derivative of the magnetic flux crossing all possible surfaces bounded by a given closed path C in item 2 implies that the time derivative of the magnetic flux emanating from a closed surface S is zero, that is,

$$\frac{d}{dt}\oint_S \mathbf{B}\cdot d\mathbf{S} = 0 \tag{1.24}$$

   One can argue then that the magnetic flux emanating from a closed surface is zero, since at an instant of time when no sources are present the magnetic field vanishes. Thus, Gauss' law for the magnetic field is not independent of Faraday's law.
9. Similarly, combining the observation concerning the sum of the current due to flow of charges and the displacement current through all possible surfaces

© 2006 by Taylor & Francis Group, LLC

bounded by a given closed path C in item 2 with the law of conservation of charge, we obtain for any closed surface S,

$$\frac{d}{dt}\left(\oint_S \mathbf{D}\cdot d\mathbf{S} - \int_V \rho\, dv\right) = 0 \tag{1.25}$$

where V is the volume bounded by S. Once again, one can then argue that the quantity inside the parentheses is zero, since at an instant of time when no sources are present, it vanishes. Thus, Gauss' law for the electric field is not independent of Ampere's circuital law in view of the law of conservation of charge.

10. The cut view in Fig. 1.8 indicates that magnetic field lines are continuous, having no beginnings or endings, whereas the cut view in Fig. 1.7 indicates that electric field lines are discontinuous wherever there are charges, diverging from positive charges and converging on negative charges.

1.2.2 Maxwell's Equations in Differential Form and the Continuity Equation

From the integral forms of Maxwell's equations, one can obtain the corresponding differential forms through the use of Stoke's and divergence theorems in vector calculus, given, respectively, by

$$\oint_C \mathbf{A}\cdot d\mathbf{l} = \int_S (\nabla \times \mathbf{A})\cdot d\mathbf{S} \tag{1.26a}$$

$$\oint_S \mathbf{A}\cdot d\mathbf{S} = \int_V (\nabla\cdot\mathbf{A})\, dv \tag{1.26b}$$

where in Eq. (1.26a), S is any surface bounded by C and in Eq. (1.26b), V is the volume bounded by S. Thus, Maxwell's equations in differential form are given by

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{1.27}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{1.28}$$

$$\nabla\cdot\mathbf{D} = \rho \tag{1.29}$$

$$\nabla\cdot\mathbf{B} = 0 \tag{1.30}$$

corresponding to the integral forms Eqs. (1.18)–(1.21), respectively. These differential equations state that at any point in a given medium, the curl of the electric field intensity is equal to the time rate of decrease of the magnetic flux density, and the curl of the magnetic field intensity is equal to the sum of the current density due to flow of charges and the displacement current density (time derivative of the displacement flux density); whereas

© 2006 by Taylor & Francis Group, LLC

the divergence of the displacement flux density is equal to the volume charge density, and the divergence of the magnetic flux density is equal to zero.

Auxiliary to the Maxwell's equations in differential form is the differential equation following from the law of conservation of charge Eq. (1.22) through the use of Eq. (1.26b). Familiarly known as the *continuity equation*, this is given by

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 \tag{1.31}$$

It states that at any point in a given medium, the divergence of the current density due to flow of charges plus the time rate of increase of the volume charge density is equal to zero.

From the interdependence of the integral laws discussed in the previous section, it follows that Eq. (1.30) is not independent of Eq. (1.27), and Eq. (1.29) is not independent of Eq. (1.28) in view of Eq. (1.31).

Maxwell's equations in differential form lend themselves well for a qualitative discussion of the interdependence of time-varying electric and magnetic fields giving rise to the phenomenon of electromagnetic wave propagation. Recognizing that the operations of curl and divergence involve partial derivatives with respect to space coordinates, we observe that time-varying electric and magnetic fields coexist in space, with the spatial variation of the electric field governed by the temporal variation of the magnetic field in accordance with Eq. (1.27), and the spatial variation of the magnetic field governed by the temporal variation of the electric field in addition to the current density in accordance with Eq. (1.28). Thus, if in Eq. (1.28) we begin with a time-varying current source represented by $\mathbf{J}$, or a time-varying electric field represented by $\partial \mathbf{D}/dt$, or a combination of the two, then one can visualize that a magnetic field is generated in accordance with Eq. (1.28), which in turn generates an electric field in accordance with Eq. (1.27), which in turn contributes to the generation of the magnetic field in accordance with Eq. (1.28), and so on, as depicted in Fig. 1.11. Note that $\mathbf{J}$ and ρ are coupled, since they must satisfy Eq. (1.31). Also, the magnetic field automatically satisfies Eq. (1.30), since Eq. (1.30) is not independent of Eq. (1.27).

The process depicted in Fig. 1.11 is exactly the phenomenon of electromagnetic waves propagating with a velocity (and other characteristics) determined by the parameters of the medium. In free space, the waves propagate unattenuated with the velocity $1/\sqrt{\mu_0 \varepsilon_0}$, familiarly represented by the symbol c. If either the term $\partial \mathbf{B}/\partial t$ in Eq. (1.27) or the term $\partial \mathbf{D}/\partial t$ in Eq. (1.28) is not present, then wave propagation would not occur. As already stated in the previous section, it was through the addition of the term

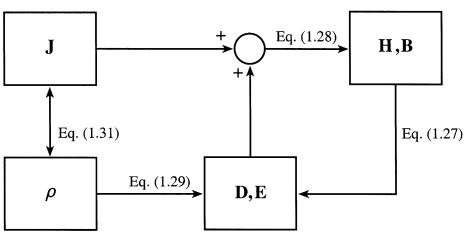


Figure 1.11 Generation of interdependent electric and magnetic fields, beginning with sources $\mathbf{J}$ and ρ.

© 2006 by Taylor & Francis Group, LLC

$\partial \mathbf{D}/\partial t$ in Eq. (1.28) that Maxwell predicted electromagnetic wave propagation before it was confirmed experimentally.

Of particular importance is the case of time variations of the fields in the sinusoidal steady state, that is, the frequency domain case. In this connection, the frequency domain forms of Maxwell's equations are of interest. Using the phasor notation based on

$$A\ \cos(\omega t + \phi) = \mathrm{Re}\left[Ae^{j\phi}e^{j\omega t}\right] = \mathrm{Re}\left[\bar{A}e^{j\omega t}\right] \tag{1.32}$$

where $\bar{A} = Ae^{j\phi}$ is the phasor corresponding to the time function, we obtain these equations by replacing all field quantities in the time domain form of the equations by the corresponding phasor quantities and $\partial/\partial t$ by $j\omega$. Thus with the understanding that all phasor field quantities are functions of space coordinates, that is, $\bar{\mathbf{E}} = \bar{\mathbf{E}}(\mathbf{r})$, etc., we write the Maxwell's equations in frequency domain as

$$\nabla \times \bar{\mathbf{E}} = -j\omega\bar{\mathbf{B}} \tag{1.33}$$

$$\nabla \times \bar{\mathbf{H}} = \bar{\mathbf{J}} + j\omega\bar{\mathbf{D}} \tag{1.34}$$

$$\nabla \cdot \bar{\mathbf{D}} = \bar{\rho} \tag{1.35}$$

$$\nabla \cdot \bar{\mathbf{B}} = 0 \tag{1.36}$$

Also, the continuity equation, Eq. (1.31), transforms to the frequency domain form

$$\nabla \cdot \bar{\mathbf{J}} + j\omega\bar{\rho} = 0 \tag{1.37}$$

Note that since $\nabla \cdot \nabla \times \bar{\mathbf{E}} = 0$, Eq. (1.36) follows from Eq. (1.33), and since $\nabla \cdot \nabla \times \bar{\mathbf{H}} = 0$, Eq. (1.35) follows from Eq. (1.34) with the aid of Eq. (1.37).

Now the constitutive relations in phasor form are

$$\bar{\mathbf{D}} = \varepsilon\bar{\mathbf{E}} \tag{1.38a}$$

$$\bar{\mathbf{H}} = \frac{\bar{\mathbf{B}}}{\mu} \tag{1.38b}$$

$$\bar{\mathbf{J}}_c = \sigma\bar{\mathbf{E}} \tag{1.38c}$$

Substituting these into Eqs. (1.33)–(1.36), we obtain for a material medium characterized by the parameters ε, μ, and σ,

$$\nabla \times \bar{\mathbf{E}} = -j\omega\mu\bar{\mathbf{H}} \tag{1.39}$$

$$\nabla \times \bar{\mathbf{H}} = (\sigma + j\omega\varepsilon)\bar{\mathbf{E}} \tag{1.40}$$

$$\nabla \cdot \bar{\mathbf{H}} = 0 \tag{1.41}$$

$$\nabla \cdot \bar{\mathbf{E}} = \frac{\bar{\rho}}{\varepsilon} \tag{1.42}$$

© 2006 by Taylor & Francis Group, LLC

Note however that if the medium is homogeneous, that is, if the material parameters are independent of the space coordinates, Eq. (1.40) gives

$$\nabla \cdot \bar{\mathbf{E}} = \frac{1}{\sigma + j\omega\varepsilon} \nabla \cdot \nabla \times \bar{\mathbf{H}} = 0 \tag{1.43}$$

so that $\bar{\rho} = 0$ in such a medium.

A point of importance in connection with the frequency domain form of Maxwell's equations is that in these equations, the parameters ε, μ, and σ can be allowed to be functions of ω. In fact, for many dielectrics, the conductivity increases with frequency in such a manner that the quantity $\sigma/\omega\varepsilon$ is more constant than is the conductivity. This quantity is the ratio of the magnitudes of the two terms on the right side of Eq. (1.40), that is, the conduction current density term $\sigma\bar{\mathbf{E}}$ and the displacement current density term $j\omega\varepsilon\bar{\mathbf{E}}$.

1.2.3. Boundary Conditions

Maxwell's equations in differential form govern the interrelationships between the field vectors and the associated source densities at points in a given medium. For a problem involving two or more different media, the differential equations pertaining to each medium provide solutions for the fields that satisfy the characteristics of that medium. These solutions need to be matched at the boundaries between the media by employing "boundary conditions," which relate the field components at points adjacent to and on one side of a boundary to the field components at points adjacent to and on the other side of that boundary. The boundary conditions arise from the fact that the integral equations involve closed paths and surfaces and they must be satisfied for all possible closed paths and surfaces whether they lie entirely in one medium or encompass a portion of the boundary.

The boundary conditions are obtained by considering one integral equation at a time and applying it to a closed path or a closed surface encompassing the boundary, as shown in Fig. 1.12 for a plane boundary, and in the limit that the area enclosed by the closed path, or the volume bounded by the closed surface, goes to zero. Let the quantities pertinent to medium 1 be denoted by subscript 1 and the quantities pertinent to medium 2 be denoted by subscript 2, and $\mathbf{a}_n$ be the unit normal vector to the surface and directed into medium 1. Let all normal components at the boundary in both media be directed along $\mathbf{a}_n$ and denoted by an additional subscript n and all tangential components at the boundary in both media be denoted by an additional subscript t. Let the surface charge density ($\mathrm{C/m^2}$) and the surface current density (A/m) on the boundary be ρ_S and $\mathbf{J}_S$, respectively. Then, the boundary conditions corresponding to the Maxwell's equations in integral form can be summarized as

$$\mathbf{a}_n \times (\mathbf{E}_1 - \mathbf{E}_2) = \mathbf{0} \tag{1.44a}$$

$$\mathbf{a}_n \times (\mathbf{H}_1 - \mathbf{H}_2) = \mathbf{J}_S \tag{1.44b}$$

$$\mathbf{a}_n \cdot (\mathbf{D}_1 - \mathbf{D}_2) = \rho_S \tag{1.44c}$$

$$\mathbf{a}_n \cdot (\mathbf{B}_1 - \mathbf{B}_2) = 0 \tag{1.44d}$$

© 2006 by Taylor & Francis Group, LLC

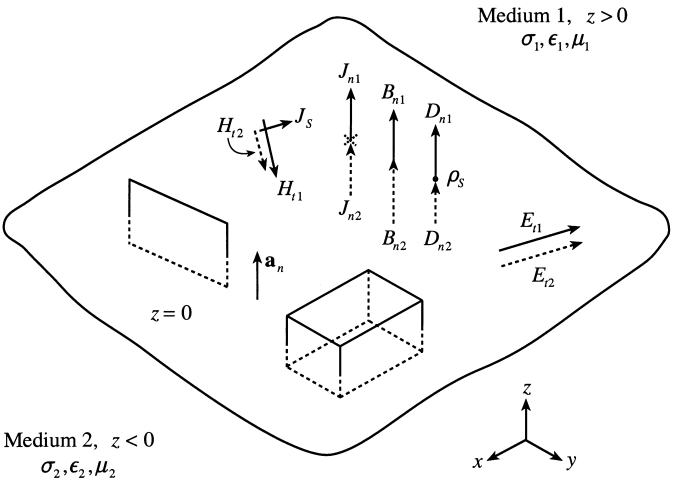


Figure 1.12 For deriving the boundary conditions at the interface between two arbitrary media.

or in scalar form,

$$E_{t1} - E_{t2} = 0 \tag{1.45a}$$
$$H_{t1} - H_{t2} = J_S \tag{1.45b}$$
$$D_{n1} - D_{n2} = \rho_S \tag{1.45c}$$
$$B_{n1} - B_{n2} = 0 \tag{1.45d}$$

In words, the boundary conditions state that at a point on the boundary, the tangential components of **E** and the normal components of **B** are continuous, whereas the tangential components of **H** are discontinuous by the amount equal to J_S at that point, and the normal components of **D** are discontinuous by the amount equal to ρ_S at that point, as illustrated in Fig. 1.12. It should be noted that the information concerning the direction of $\mathbf{J}_S$ relative to that of $(\mathbf{H}_1 - \mathbf{H}_2)$, which is contained in Eq. (1.44b), is not present in Eq. (1.45b). Hence, in general, Eq. (1.45b) is not sufficient and it is necessary to use Eq. (1.44b).

While Eqs. (1.44a)–(1.44d) or Eqs. (1.45a)–(1.45d) are the most commonly used boundary conditions, another useful boundary condition resulting from the law of conservation of charge is given by

$$\mathbf{a}_n \cdot (\mathbf{J}_1 - \mathbf{J}_2) = -\nabla_S \cdot \mathbf{J}_S - \frac{\partial \rho_S}{\partial t} \tag{1.46}$$

In words, Eq. (1.46) states that, at any point on the boundary, the components of $\mathbf{J}_1$ and $\mathbf{J}_2$ normal to the boundary are discontinuous by the amount equal to the negative of the sum of the two-dimensional divergence of the surface current density and the time derivative of the surface charge density at that point.

© 2006 by Taylor & Francis Group, LLC

1.2.4. Electromagnetic Potentials and Potential Function Equations

Maxwell's equations in differential form, together with the constitutive relations and boundary conditions, allow for the unique determination of the fields **E**, **B**, **D**, and **H** for a given set of source distributions with densities **J** and ρ. An alternate approach involving the electric scalar potential Φ and the magnetic vector potential **A**, known together as the *electromagnetic potentials* from which the fields can be derived, simplifies the solution in some cases. This approach leads to solving two separate differential equations, one for Φ involving ρ alone, and the second for **A** involving **J** alone.

To obtain these equations, we first note that in view of Eq. (1.30), **B** can be expressed as the curl of another vector. Thus

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{1.47}$$

Note that the units of **A** are the units of **B** times meter, that is, Wb/m. Now, substituting Eq. (1.47) into Eq. (1.27), interchanging the operations of $\partial/\partial t$ and curl, and rearranging, we obtain

$$\begin{aligned} \nabla \times \left[\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}\right] &= 0 \\ \mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} &= -\nabla \Phi \\ \mathbf{E} &= -\nabla \Phi - \frac{\partial \mathbf{A}}{\partial t} \end{aligned} \tag{1.48}$$

where the negative sign associated with $\nabla\Phi$ is chosen for a reason to be evident later in Sec. 1.3.2. Note that the units of Φ are the units of **E** times meter, that is, V. Note also that the knowledge of Φ and **A** enables the determination of **E** and **B**, from which **D** and **H** can be found by using the constitutive relations.

Now, using Eqs. (1.6) and (1.10) to obtain **D** and **H** in terms of Φ and **A** and substituting into Eqs. (1.29) and (1.28), we obtain

$$\nabla^2 \Phi + \nabla \cdot \left[\frac{\partial \mathbf{A}}{\partial t}\right] = -\frac{\rho}{\varepsilon} \tag{1.49a}$$

$$\nabla \times \nabla \times \mathbf{A} + \mu\varepsilon \frac{\partial}{\partial t}\left[\nabla \Phi + \frac{\partial \mathbf{A}}{\partial t}\right] = \mu \mathbf{J} \tag{1.49b}$$

where we have assumed the medium to be homogeneous and isotropic, in addition to being linear. Using the vector identity

$$\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} \tag{1.50}$$

and interchanging the operations of $\partial/\partial t$ and divergence or gradient depending on the term, and rearranging, we get

$$\nabla^2 \Phi + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) = -\frac{\rho}{\varepsilon} \tag{1.51a}$$

© 2006 by Taylor & Francis Group, LLC

$$\nabla^2\mathbf{A} - \nabla\left[\nabla\cdot\mathbf{A} + \mu\varepsilon\frac{\partial\Phi}{\partial t}\right] - \mu\varepsilon\ \frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\mathbf{J} \tag{1.51b}$$

These equations are coupled. To uncouple them, we make use of Helmholtz's theorem, which states that a vector field is completely specified by its curl and divergence. Therefore, since the curl of **A** is given by Eq. (1.47), we are at liberty to specify the divergence of **A**. We do this by setting

$$\nabla\cdot\mathbf{A} = -\mu\varepsilon\ \frac{\partial\Phi}{\partial t} \tag{1.52}$$

which is known as the Lorenz condition, resulting in the uncoupled equations

$$\nabla^2\Phi - \mu\varepsilon\ \frac{\partial^2\Phi}{\partial t^2} = -\frac{\rho}{\varepsilon} \tag{1.53}$$

$$\nabla^2\mathbf{A} - \mu\varepsilon\ \frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\mathbf{J} \tag{1.54}$$

which are called the *potential function equations*. While the Lorenz condition may appear to be arbitrary, it actually implies the continuity equation, which can be shown by taking the Laplacian on both sides of Eq. (1.52) and using Eqs. (1.53) and (1.54).

It can be seen that Eqs. (1.53) and (1.54) are not only uncoupled but they are also similar, particularly in Cartesian coordinates since Eq. (1.54) decomposes into three equations involving the three Cartesian components of **J**, each of which is similar to (1.53). By solving Eqs. (1.53) and (1.54), one can obtain the solutions for Φ and **A**, respectively, from which **E** and **B** can be found by using Eqs. (1.48) and (1.47), respectively. In practice, however, since ρ is related to **J** through the continuity equation, it is sufficient to find **B** from **A** obtained from the solution of Eq. (1.54) and then find **E** by using the Maxwell's equation for the curl of **H**, given by Eq. (1.28).

1.2.5. Power Flow and Energy Storage

A unique property of the electromagnetic field is its ability to transfer power between two points even in the absence of an intervening material medium. Without such ability, the effect of the field generated at one point will not be felt at another point, and hence the power generated at the first point cannot be put to use at the second point.

To discuss power flow associated with an electromagnetic field, we begin with the vector identity

$$\nabla\cdot(\mathbf{E}\times\mathbf{H}) = \mathbf{H}\cdot(\nabla\times\mathbf{E}) - \mathbf{E}\cdot(\nabla\times\mathbf{H}) \tag{1.55}$$

and make use of Maxwell's curl equations, Eqs. (1.27) and (1.28), to write

$$\nabla\cdot(\mathbf{E}\times\mathbf{H}) = -\mathbf{E}\cdot\mathbf{J} - \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t} - \mathbf{H}\cdot\frac{\partial\mathbf{B}}{\partial t} \tag{1.56}$$

Allowing for conductivity of a material medium by denoting $\mathbf{J} = \mathbf{J}_0 + \mathbf{J}_c$, where $\mathbf{J}_0$ is that part of **J** that can be attributed to a source, and using the constitutive relations (1.5), (1.6),

© 2006 by Taylor & Francis Group, LLC

and (1.10), we obtain for a medium characterized by σ, ε, and μ,

$$-\mathbf{E}\cdot\mathbf{J}_0 = \sigma\varepsilon^2 + \frac{\partial}{\partial t}\left[\frac{1}{2}\varepsilon E^2\right] + \frac{\partial}{\partial t}\left[\frac{1}{2}\mu H^2\right] + \nabla\cdot(\mathbf{E}\times\mathbf{H}) \tag{1.57}$$

Defining a vector $\mathbf{P}$ given by

$$\mathbf{P} = \mathbf{E}\times\mathbf{H} \tag{1.58}$$

and taking the volume integral of both sides of Eq. (1.58), we obtain

$$-\int_V (\mathbf{E}\cdot\mathbf{J}_0)\, dv = \int_V \sigma E^2\, dv + \frac{\partial}{\partial t}\int_V \left(\frac{1}{2}\varepsilon E^2\right) dv + \frac{\partial}{\partial t}\int_V \left(\frac{1}{2}\mu H^2\right) dv + \oint_S \mathbf{P}\cdot d\mathbf{S} \tag{1.59}$$

where we have also interchanged the differentiation operation with time and integration operation over volume in the second and third terms on the right side and used the divergence theorem for the last term.

In Eq. (1.59), the left side is the power supplied to the field by the current source $\mathbf{J}_0$ inside V. The quantities σE^2, $(1/2)\varepsilon E^2$, and $(1/2)\mu H^2$ are the power dissipation density ($\mathrm{W/m^3}$), the electric stored energy density ($\mathrm{J/m^3}$), and the magnetic stored energy density ($\mathrm{J/m^3}$), respectively, due to the conductive, dielectric, and magnetic properties, respectively, of the medium. Hence, Eq. (1.59) says that the power delivered to the volume V by the current source $\mathbf{J}_0$ is accounted for by the power dissipated in the volume due to the conduction current in the medium, plus the time rates of increase of the energies stored in the electric and magnetic fields, plus another term, which we must interpret as the power carried by the electromagnetic field out of the volume V, for conservation of energy to be satisfied. It then follows that the vector $\mathbf{P}$ has the meaning of power flow density vector associated with the electromagnetic field. The statement represented by Eq. (1.59) is known as the *Poynting's theorem*, and the vector $\mathbf{P}$ is known as the *Poynting vector*. We note that the units of $\mathbf{E}\times\mathbf{H}$ are volts per meter times amperes per meter, or watts per square meter ($\mathrm{W/m^2}$) and do indeed represent power density. In particular, since $\mathbf{E}$ and $\mathbf{H}$ are instantaneous field vectors, $\mathbf{E}\times\mathbf{H}$ represents the instantaneous Poynting vector. Note that the Poynting's theorem tells us only that the power flow out of a volume V is given by the surface integral of the Poynting vector over the surface S bounding that volume. Hence we can add to $\mathbf{P}$ any vector for which the surface integral over S vanishes, without affecting the value of the surface integral. However, generally, we are interested in the total power leaving a closed surface and the interpretation of $\mathbf{P}$ alone as representing the power flow density vector is sufficient.

For sinusoidally time-varying fields, that is, for the frequency domain case, the quantity of importance is the time-average Poynting vector instead of the instantaneous Poynting vector. We simply present the important relations here, without carrying out the derivations. The time-average Poynting vector, denoted by $\langle\mathbf{P}\rangle$, is given by

$$\langle\mathbf{P}\rangle = \mathrm{Re}\left[\bar{\mathbf{P}}\right] \tag{1.60}$$

© 2006 by Taylor & Francis Group, LLC

where $\bar{\mathbf{P}}$ is the complex Poynting vector given by

$$\bar{\mathbf{P}} = \frac{1}{2}\bar{\mathbf{E}} \times \bar{\mathbf{H}}^* \tag{1.61}$$

where the star denotes complex conjugate. The Poynting theorem for the frequency domain case, known as the *complex Poynting's theorem*, is given by

$$-\int_V \left(\frac{1}{2}\bar{\mathbf{E}}\cdot\bar{\mathbf{J}}_0^*\right) dv = \int_V \langle p_d\rangle dv + j2\omega \int_V (\langle w_m\rangle - \langle w_e\rangle) dv + \oint_S \bar{\mathbf{P}}\cdot d\mathbf{S} \tag{1.62}$$

where

$$\langle p_d\rangle = \frac{1}{2}\sigma\bar{\mathbf{E}}\cdot\bar{\mathbf{E}}^* \tag{1.63a}$$

$$\langle w_e\rangle = \frac{1}{4}\varepsilon\bar{\mathbf{E}}\cdot\bar{\mathbf{E}}^* \tag{1.63b}$$

$$\langle w_m\rangle = \frac{1}{4}\mu\bar{\mathbf{H}}\cdot\bar{\mathbf{H}}^* \tag{1.63c}$$

are the time-average power dissipation density, the time-average electric stored energy density, and the time-average magnetic stored energy density, respectively. Equation (1.62) states that the time-average, or real, power delivered to the volume V by the current source is accounted for by the time-average power dissipated in the volume plus the time-average power carried by the electromagnetic field out of the volume through the surface S bounding the volume and that the reactive power delivered to the volume V by the current source is equal to the reactive power carried by the electromagnetic field out of the volume V through the surface S plus a quantity that is 2ω times the difference between the time-average magnetic and electric stored energies in the volume.

1.3. STATIC FIELDS, QUASISTATIC FIELDS, AND WAVES

1.3.1. Classification of Fields

While every macroscopic field obeys Maxwell's equations in their entirety, depending on their most dominant properties, it is sufficient to consider a subset of, or certain terms only, in the equations. The primary classification of fields is based on their time dependence. Fields which do not change with time are called *static*. Fields which change with time are called *dynamic*. Static fields are the simplest kind of fields, because for them $\partial/\partial t = 0$ and all terms involving differentiation with respect to time go to zero. Dynamic fields are the most complex, since for them Maxwell's equations in their entirety must be satisfied, resulting in wave type solutions, as provided by the qualitative explanation in Sec. 1.2.2. However, if certain features of the dynamic field can be analyzed as though the field were static, then the field is called *quasistatic*.

If the important features of the field are not amenable to static type field analysis, they are generally referred to as *time-varying*, although in fact, quasistatic fields are also time-varying. Since in the most general case, time-varying fields give rise to wave phenomena, involving velocity of propagation and time delay, it can be said that quasistatic fields are those time-varying fields for which wave propagation effects can be neglected.

© 2006 by Taylor & Francis Group, LLC

1.3.2. Static Fields and Circuit Elements

For static fields, $\partial/\partial t = 0$. Maxwell's equations in integral form and the law of conservation of charge become

$$\oint_C \mathbf{E}\cdot d\mathbf{l} = 0 \tag{1.64a}$$

$$\oint_C \mathbf{H}\cdot d\mathbf{l} = \int_S \mathbf{J}\cdot d\mathbf{S} \tag{1.64b}$$

$$\oint_S \mathbf{D}\cdot d\mathbf{S} = \int_V \rho\, dv \tag{1.64c}$$

$$\oint_S \mathbf{B}\cdot d\mathbf{S} = 0 \tag{1.64d}$$

$$\oint_S \mathbf{J}\cdot d\mathbf{S} = 0 \tag{1.64e}$$

whereas Maxwell's equations in differential form and the continuity equation reduce to

$$\nabla \times \mathbf{E} = 0 \tag{1.65a}$$

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{1.65b}$$

$$\nabla\cdot\mathbf{D} = \rho \tag{1.65c}$$

$$\nabla\cdot\mathbf{B} = 0 \tag{1.65d}$$

$$\nabla\cdot\mathbf{J} = 0 \tag{1.65e}$$

Immediately, one can see that, unless $\mathbf{J}$ includes a component due to conduction current, the equations involving the electric field are completely independent of those involving the magnetic field. Thus the fields can be subdivided into *static electric fields*, or *electrostatic fields*, governed by Eqs. (1.64a) and (1.64c), or Eqs. (1.65a) and (1.65c), and *static magnetic fields*, or *magnetostatic fields*, governed by Eqs. (1.64b) and (1.64d), or Eqs. (1.65b) and (1.65d). The source of a static electric field is ρ, whereas the source of a static magnetic field is $\mathbf{J}$. One can also see from Eq. (1.64e) or (1.65e) that no relationship exists between $\mathbf{J}$ and ρ. If $\mathbf{J}$ includes a component due to conduction current, then since $\mathbf{J}_c = \sigma\mathbf{E}$, a coupling between the electric and magnetic fields exists for that part of the total field associated with $\mathbf{J}_c$. However, the coupling is only one way, since the right side of Eq. (1.64a) or (1.65a) is still zero. The field is then referred to as *electromagnetostatic field*. It can also be seen then that for consistency, the right sides of Eqs. (1.64c) and (1.65c) must be zero, since the right sides of Eqs. (1.64e) and (1.65e) are zero. We shall now consider each of the three types of static fields separately and discuss some fundamental aspects.

Electrostatic Fields and Capacitance

The equations of interest are Eqs. (1.64a) and (1.64c), or Eqs. (1.65a) and (1.65c). The first of each pair of these equations simply tells us that the electrostatic field is a conservative

© 2006 by Taylor & Francis Group, LLC

field, and the second of each pair of these equations enables us, in principle, to determine the electrostatic field for a given charge distribution. Alternatively, the potential function equation, Eq. (1.53), which reduces to

$$\nabla^2 \Phi = -\frac{\rho}{\varepsilon} \tag{1.66}$$

can be used to find the electric scalar potential, Φ, from which the electrostatic field can be determined by using Eq. (1.48), which reduces to

$$\mathbf{E} = -\nabla \Phi \tag{1.67}$$

Equation (1.66) is known as the *Poisson's equation*, which automatically includes the condition that the field be conservative. It is worth noting that the potential difference between two points A and B in the static electric field, which is independent of the path followed from A to B because of the conservative nature of the field is

$$\begin{aligned} \int_A^B \mathbf{E} \cdot d\mathbf{l} &= \int_A^B [-\nabla \Phi] \cdot d\mathbf{l} \\ &= \Phi_A - \Phi_B \end{aligned} \tag{1.68}$$

the difference between the value of Φ at A and the value of Φ at B. The choice of minus sign associated with $\nabla\Phi$ in Eq. (1.48) is now evident.

The solution to Poisson's equation, Eq. (1.66), for a given charge density distribution $\rho(\mathbf{r})$ is given by

$$\Phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon} \int_{V'} \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, dv' \tag{1.69}$$

where the prime denotes source point and no prime denotes field point. Although cast in terms of volume charge density, Eq. (1.69) can be formulated in terms of a surface charge distribution, a line charge distribution, or a collection of point charges. In particular, for a point charge $Q(\mathbf{r}')$, the solution is given by

$$\Phi(\mathbf{r}) = \frac{Q(\mathbf{r}')}{4\pi\varepsilon |\mathbf{r} - \mathbf{r}'|} \tag{1.70}$$

It follows from Eq. (1.67) that the electric field intensity due to the point charge is given by

$$\mathbf{E}(\mathbf{r}) = \frac{Q(\mathbf{r}')\,(\mathbf{r} - \mathbf{r}')}{4\pi\varepsilon\,|\mathbf{r} - \mathbf{r}'|^3} \tag{1.71}$$

which is exactly the expression that results from Coulomb's law for the electric force between two point charges.

© 2006 by Taylor & Francis Group, LLC

Equation (1.69) or its alternate forms can be used to solve two types of problems:

1. finding the electrostatic potential for a specified charge distribution by evaluating the integral on the right side, which is a straightforward process with the help of a computer but can be considerably difficult analytically except for a few examples, and
2. finding the surface charge distribution on the surfaces of an arrangement of conductors raised to specified potentials, by inversion of the equation, which is the basis for numerical solution by the well-known *method of moments*.

In the case of type 1, the electric field can then be found by using Eq. (1.67).

In a charge-free region, $\rho = 0$, and Poisson's equation, Eq. (1.66), reduces to

$$\nabla^2 \Phi = 0 \tag{1.72}$$

which is known as the *Laplace equation*. The field is then due to charges outside the region, such as surface charge on conductors bounding the region. The situation is then one of solving a boundary value problem. In general, for arbitrarily shaped boundaries, a numerical technique, such as the *method of finite differences*, is employed for solving the problem. Here, we consider analytical solution involving one-dimensional variation of Φ.

A simple example is that of the parallel-plate arrangement shown in Fig. 1.13a, in which two parallel, perfectly conducting plates ($\sigma = \infty$, $\mathbf{E} = 0$) of dimensions w along the y direction and l along the z direction lie in the $x = 0$ and $x = d$ planes. The region between the plates is a perfect dielectric ($\sigma = 0$) of material parameters ε and μ. The thickness of the plates is shown exaggerated for convenience in illustration. A potential difference of V_0 is maintained between the plates by connecting a direct voltage source at the end $z = -l$. If fringing of the field due to the finite dimensions of the structure normal to the x direction is neglected, or if it is assumed that the structure is part of one which is infinite in extent

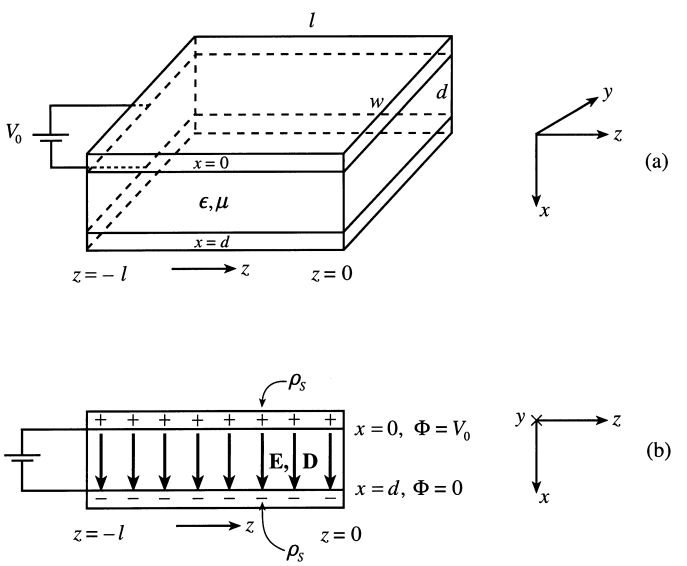


Figure 1.13 Electrostatic field in a parallel-plate arrangement.

© 2006 by Taylor & Francis Group, LLC

normal to the x direction, then the problem can be treated as one-dimensional with x as the variable, and Eq. (1.72) reduces to

$$\frac{d^2\Phi}{dx^2} = 0 \tag{1.73}$$

The solution for the potential in the charge-free region between the plates is given by

$$\Phi(x) = \frac{V_0}{d}(d - x) \tag{1.74}$$

which satisfies Eq. (1.73), as well as the boundary conditions of $\Phi = 0$ at $x = d$ and $\Phi = V_0$ at $x = 0$. The electric field intensity between the plates is then given by

$$\mathbf{E} = -\nabla\Phi = \frac{V_0}{d}\mathbf{a}_x \tag{1.75}$$

as depicted in the cross-sectional view in Fig. 1.13b, and resulting in displacement flux density

$$\mathbf{D} = \frac{\varepsilon V_0}{d}\mathbf{a}_x \tag{1.76}$$

Then, using the boundary condition for the normal component of $\mathbf{D}$ given by Eq. (1.44c) and noting that there is no field inside the conductor, we obtain the magnitude of the charge on either plate to be

$$Q = \left(\frac{\varepsilon V_0}{d}\right)(wl) = \frac{\varepsilon wl}{d}V_0 \tag{1.77}$$

We can now find the familiar circuit parameter, the capacitance, C, of the parallel-plate arrangement, which is defined as the ratio of the magnitude of the charge on either plate to the potential difference V_0. Thus

$$C = \frac{Q}{V_0} = \frac{\varepsilon wl}{d} \tag{1.78}$$

Note that the units of C are the units of ε times meter, that is, farads. The phenomenon associated with the arrangement is that energy is stored in the capacitor in the form of electric field energy between the plates, as given by

$$\begin{aligned} W_e &= \left(\frac{1}{2}\varepsilon E_x^2\right)(wld) \\ &= \frac{1}{2}\left(\frac{\varepsilon wl}{d}\right)V_0^2 \\ &= \frac{1}{2}CV_0^2 \end{aligned} \tag{1.79}$$

the familiar expression for energy stored in a capacitor.

© 2006 by Taylor & Francis Group, LLC

Magnetostatic Fields and Inductance

The equations of interest are Eqs. (1.64b) and (1.64d) or Eqs. (1.65b) and (1.65d). The second of each pair of these equations simply tells us that the magnetostatic field is solenoidal, which as we know holds for any magnetic field, and the first of each pair of these equations enables us, in principle, to determine the magnetostatic field for a given current distribution. Alternatively, the potential function equation, Eq. (1.54), which reduces to

$$\nabla^2 \mathbf{A} = -\mu \mathbf{J} \tag{1.80}$$

can be used to find the magnetic vector potential, **A**, from which the magnetostatic field can be determined by using Eq. (1.47). Equation (1.80) is the Poisson's equation for the magnetic vector potential, which automatically includes the condition that the field be solenoidal.

The solution to Eq. (1.80) for a given current density distribution $\mathbf{J}(\mathbf{r})$ is, purely from analogy with the solution Eq. (1.69) to Eq. (1.66), given by

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi} \int_{V'} \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, dv' \tag{1.81}$$

Although cast in terms of volume current density, Eq. (1.81) can be formulated in terms of a surface current density, a line current, or a collection of infinitesimal current elements. In particular, for an infinitesimal current element $I\,d\mathbf{l}(\mathbf{r}')$, the solution is given by

$$\mathbf{A}(\mathbf{r}) = \frac{\mu I \, d\mathbf{l}(\mathbf{r}')}{4\pi |\mathbf{r} - \mathbf{r}'|} \tag{1.82}$$

It follows from Eq. (1.47) that the magnetic flux density due to the infinitesimal current element is given by

$$\mathbf{B}(\mathbf{r}) = \frac{\mu I \, d\mathbf{l}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}')}{4\pi |\mathbf{r} - \mathbf{r}'|^3} \tag{1.83}$$

which is exactly the law of Biot-Savart that results from Ampere's force law for the magnetic force between two current elements. Similar to that in the case of Eq. (1.69), Eq. (1.81) or its alternate forms can be used to find the magnetic vector potential and then the magnetic field by using Eq. (1.47) for a specified current distribution.

In a current-free region, $\mathbf{J} = 0$, and Eq. (1.80) reduces to

$$\nabla^2 \mathbf{A} = 0 \tag{1.84}$$

The field is then due to currents outside the region, such as surface currents on conductors bounding the region. The situation is then one of solving a boundary value problem as in the case of Eq. (1.72). However, since the boundary condition Eq. (1.44b) relates the magnetic field directly to the surface current density, it is straightforward and more convenient to determine the magnetic field directly by using Eqs. (1.65b) and (1.65d).

© 2006 by Taylor & Francis Group, LLC

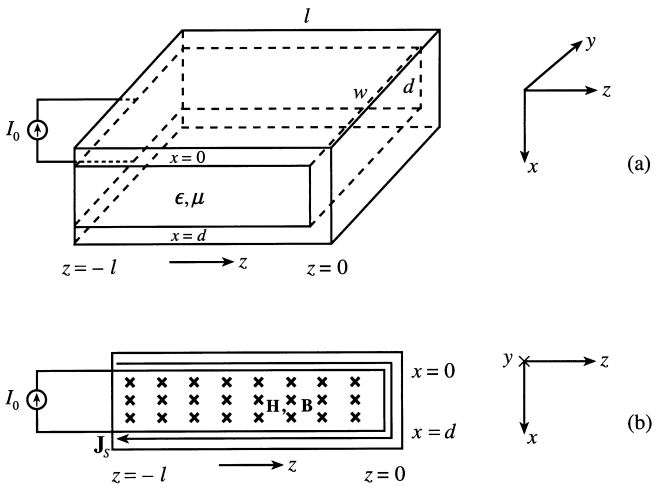


Figure 1.14 Magnetostatic field in a parallel-plate arrangement.

A simple example is that of the parallel-plate arrangement of Fig. 1.13a with the plates connected by another conductor at the end $z = 0$ and driven by a source of direct current I_0 at the end $z = -l$, as shown in Fig. 1.14a. If fringing of the field due to the finite dimensions of the structure normal to the x direction is neglected, or if it is assumed that the structure is part of one which is infinite in extent normal to the x direction, then the problem can be treated as one-dimensional with x as the variable and we can write the current density on the plates to be

$$\mathbf{J}_S = \begin{cases} \left(\dfrac{I_0}{w}\right)\mathbf{a}_z & \text{on the plate } x = 0 \\ \left(\dfrac{I_0}{w}\right)\mathbf{a}_x & \text{on the plate } z = 0 \\ -\left(\dfrac{I_0}{w}\right)\mathbf{a}_z & \text{on the plate } x = d \end{cases} \tag{1.85}$$

In the current-free region between the plates, Eq. (1.65b) reduces to

$$\begin{vmatrix} \mathbf{a}_x & \mathbf{a}_y & \mathbf{a}_z \\ \dfrac{\partial}{\partial x} & 0 & 0 \\ H_x & H_y & H_z \end{vmatrix} = 0 \tag{1.86}$$

and Eq. (1.65d) reduces to

$$\frac{\partial B_x}{\partial x} = 0 \tag{1.87}$$

so that each component of the field, if it exists, has to be uniform. This automatically forces H_x and H_z to be zero since nonzero value of these components do not satisfy the boundary conditions Eqs. (1.44b) and (1.44d) on the plates, keeping in mind that the field

© 2006 by Taylor & Francis Group, LLC

is entirely in the region between the conductors. Thus, as depicted in the cross-sectional view in Fig. 1.14b,

$$\mathbf{H} = \frac{I_0}{w}\mathbf{a}_y \tag{1.88}$$

which satisfies the boundary condition Eq. (1.44b) on all three plates, and results in magnetic flux density

$$\mathbf{B} = \frac{\mu I_0}{w}\mathbf{a}_y \tag{1.89}$$

The magnetic flux, ψ, linking the current I_0, is then given by

$$\psi = \left(\frac{\mu I_0}{w}\right)(dl) = \frac{\mu dl}{w} I_0 \tag{1.90}$$

We can now find the familiar circuit parameter, the inductance, L, of the parallel-plate arrangement, which is defined as the ratio of the magnetic flux linking the current to the current. Thus

$$L = \frac{\psi}{I_0} = \frac{\mu dl}{w} \tag{1.91}$$

Note that the units of L are the units of μ times meter, that is, henrys. The phenomenon associated with the arrangement is that energy is stored in the inductor in the form of magnetic field energy between the plates, as given by

$$\begin{aligned} W_m &= \left(\frac{1}{2}\mu H^2\right) wld \\ &= \frac{1}{2}\left(\frac{\mu dl}{w}\right) I_0^2 \\ &= \frac{1}{2} L I_0^2 \end{aligned} \tag{1.92}$$

the familiar expression for energy stored in an inductor.

Electromagnetostatic Fields and Conductance

The equations of interest are

$$\oint_C \mathbf{E}\cdot d\mathbf{l} = 0 \tag{1.93a}$$

$$\oint_C \mathbf{H}\cdot d\mathbf{l} = \int_S \mathbf{J}_c \cdot d\mathbf{S} = \sigma \int_S \mathbf{E}\cdot d\mathbf{S} \tag{1.93b}$$

$$\oint_S \mathbf{D}\cdot d\mathbf{S} = 0 \tag{1.93c}$$

© 2006 by Taylor & Francis Group, LLC

$$\oint_S \mathbf{B}\cdot d\mathbf{S} = 0 \tag{1.93d}$$

or in differential form,

$$\nabla \times \mathbf{E} = 0 \tag{1.94a}$$

$$\nabla \times \mathbf{H} = \mathbf{J}_c = \sigma \mathbf{E} \tag{1.94b}$$

$$\nabla \cdot \mathbf{D} = 0 \tag{1.94c}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{1.94d}$$

From Eqs. (1.94a) and (1.94c), we note that Laplace's equation, Eq. (1.72), for the electrostatic potential is satisfied, so that, for a given problem, the electric field can be found in the same manner as in the case of the example of Fig. 1.13. The magnetic field is then found by using Eq. (1.94b) and making sure that Eq. (1.94d) is also satisfied.

A simple example is that of the parallel-plate arrangement of Fig. 1.13a but with an imperfect dielectric material of parameters σ, ε, and μ, between the plates, as shown in Fig. 1.15a. Then, the electric field between the plates is the same as that given by Eq. (1.75), that is,

$$\mathbf{E} = \frac{V_0}{d}\mathbf{a}_x \tag{1.95}$$

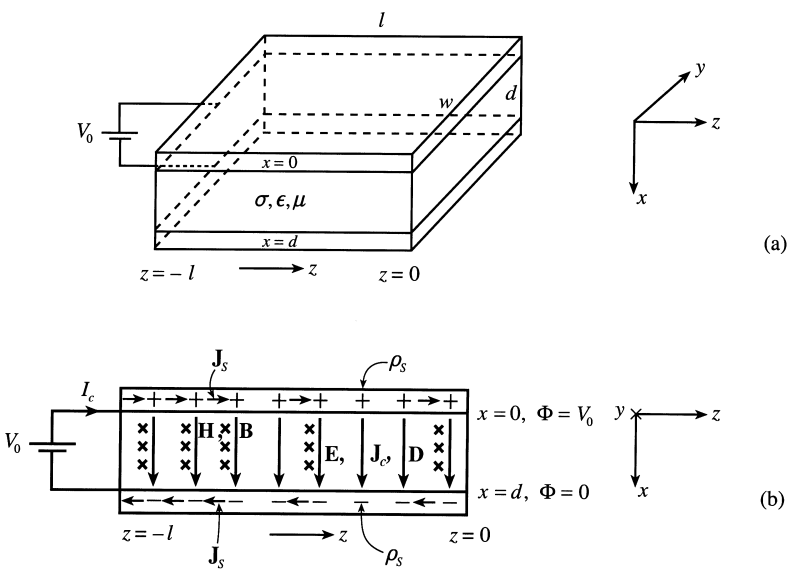


Figure 1.15 Electromagnetostatic field in a parallel-plate arrangement.

© 2006 by Taylor & Francis Group, LLC

resulting in a conduction current of density

$$\mathbf{J}_c = \frac{\sigma V_0}{d}\mathbf{a}_x \tag{1.96}$$

from the top plate to the bottom plate, as depicted in the cross-sectional view of Fig. 1.15b. Since $\partial\rho/\partial t = 0$ at the boundaries between the plates and the slab, continuity of current is satisfied by the flow of surface current on the plates. At the input $z = -l$, this surface current, which is the current drawn from the source, must be equal to the total current flowing from the top to the bottom plate. It is given by

$$I_c = \left(\frac{\sigma V_0}{d}\right)(wl) = \frac{\sigma wl}{d}V_0 \tag{1.97}$$

We can now find the familiar circuit parameter, the conductance, G, of the parallel-plate arrangement, which is defined as the ratio of the current drawn from the source to the source voltage V_0. Thus

$$G = \frac{I_c}{V_0} = \frac{\sigma wl}{d} \tag{1.98}$$

Note that the units of G are the units of σ times meter, that is, siemens (S). The reciprocal quantity, R, the resistance of the parallel-plate arrangement, is given by

$$R = \frac{V_0}{I_c} = \frac{d}{\sigma wl} \tag{1.99}$$

The unit of R is ohms. The phenomenon associated with the arrangement is that power is dissipated in the material between the plates, as given by

$$\begin{aligned} P_d &= (\sigma E^2)(wld) \\ &= \left(\frac{\sigma wl}{d}\right)V_0^2 \\ &= G\,V_0^2 \\ &= \frac{V_0^2}{R} \end{aligned} \tag{1.100}$$

the familiar expression for power dissipated in a resistor.

Proceeding further, we find the magnetic field between the plates by using Eq. (1.94b), and noting that the geometry of the situation requires a y component of $\mathbf{H}$, dependent on z, to satisfy the equation. Thus

$$\mathbf{H} = H_y(z)\,\mathbf{a}_y \tag{1.101a}$$

© 2006 by Taylor & Francis Group, LLC

$$\frac{\partial H_y}{\partial z} = -\frac{\sigma V_0}{d} \tag{1.101b}$$

$$\mathbf{H} = -\frac{\sigma V_0}{d} z\, \mathbf{a}_y \tag{1.101c}$$

where the constant of integration is set to zero, since the boundary condition at $z = 0$ requires H_y to be zero for z equal to zero. Note that the magnetic field is directed in the positive y direction (since z is negative) and increases linearly from $z = 0$ to $z = -l$, as depicted in Fig. 1.15b. It also satisfies the boundary condition at $z = -l$ by being consistent with the current drawn from the source to be $w[H_y]_{z=-1} = (\sigma V_0/d)(wl) = I_c$.

Because of the existence of the magnetic field, the arrangement is characterized by an inductance, which can be found either by using the flux linkage concept or by the energy method. To use the flux linkage concept, we recognize that a differential amount of magnetic flux $d\psi' = \mu H_y d(dz')$ between z equal to $(z' - dz')$ and z equal to z', where $-l < z' < 0$, links only that part of the current that flows from the top plate to the bottom plate between $z = z'$ and $z = 0$, thereby giving a value of $(-z'/l)$ for the fraction, N, of the total current linked. Thus, the inductance, familiarly known as the *internal inductance*, denoted L_i, since it is due to magnetic field internal to the current distribution, as compared to that in Eq. (1.91) for which the magnetic field is external to the current distribution, is given by

$$\begin{aligned} L_i &= \frac{1}{I_c}\int_{z'=-l}^{0} N\, d\psi' \\ &= \frac{1}{3}\frac{\mu dl}{w} \end{aligned} \tag{1.102}$$

or 1/3 times the inductance of the structure if $\sigma = 0$ and the plates are joined at $z = 0$, as in Fig. 1.14b.

Alternatively, if the energy method is used by computing the energy stored in the magnetic field and setting it equal to $(1/2)\, L_i I_c^2$, then we have

$$\begin{aligned} L_i &= \frac{1}{I_c^2}(dw)\int_{z=-l}^{0} \mu H_y^2\, dz \\ &= \frac{1}{3}\frac{\mu dl}{w} \end{aligned} \tag{1.103}$$

same as in Eq. (1.102).

Finally, recognizing that there is energy storage associated with the electric field between the plates, we note that the arrangement has also associated with it a capacitance C, equal to $\varepsilon wl/d$. Thus, all three properties of conductance, capacitance, and inductance are associated with the structure. Since for $\sigma = 0$ the situation reduces to that of Fig. 1.13, we can represent the arrangement of Fig. 1.15 to be equivalent to the circuit shown in Fig. 1.16. Note that the capacitor is charged to the voltage V_0 and the current through it is zero (open circuit condition). The voltage across the inductor is zero (short circuit condition), and the current through it is V_0/R. Thus, the current drawn from the voltage

© 2006 by Taylor & Francis Group, LLC

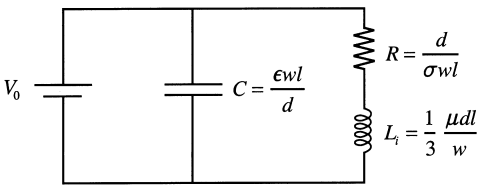


Figure 1.16 Circuit equivalent for the arrangement of Fig. 1.15.

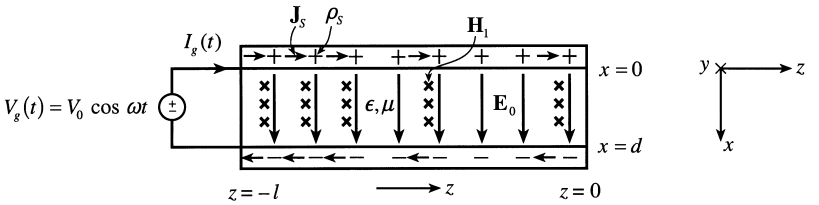


Figure 1.17 Electroquasistatic field analysis for the parallel-plate structure of Fig. 1.13.

source is V_0/R and the voltage source views a single resistor R, as far as the current drawn from it is concerned.

1.3.3. Quasistatic Fields and Low-frequency Behavior

As mentioned in Sec. 1.3.1, quasistatic fields are a class of dynamic fields for which certain features can be analyzed as though the fields were static. In terms of behavior in the frequency domain, they are low-frequency extensions of static fields present in a physical structure, when the frequency of the source driving the structure is zero, or low-frequency approximations of time-varying fields in the structure that are complete solutions to Maxwell's equations. Here, we use the approach of low-frequency extensions of static fields. Thus, for a given structure, we begin with a time-varying field having the same spatial characteristics as that of the static field solution for the structure and obtain field solutions containing terms up to and including the first power (which is the lowest power) in ω for their amplitudes. Depending on whether the predominant static field is electric or magnetic, quasistatic fields are called *electroquasistatic fields* or *magnetoquasistatic fields*. We shall now consider these separately.

Electroquasistatic Fields

For electroquasistatic fields, we begin with the electric field having the spatial dependence of the static field solution for the given arrangement. An example is provided by the arrangement in Fig. 1.13a excited by a sinusoidally time-varying voltage source $V_g(t) = V_0 \cos \omega t$, instead of a direct voltage source, as shown by the cross-sectional view in Fig. 1.17. Then,

$$\mathbf{E}_0 = \frac{V_0}{d} \cos \omega t \, \mathbf{a}_x \tag{1.104}$$

© 2006 by Taylor & Francis Group, LLC

where the subscript 0 denotes that the amplitude of the field is of the zeroth power in ω. This results in a magnetic field in accordance with Maxwell's equation for the curl of $\mathbf{H}$, given by Eq. (1.28). Thus, noting that $\mathbf{J}=0$ in view of the perfect dielectric medium, we have for the geometry of the arrangement,

$$\frac{\partial H_{y1}}{\partial z} = -\frac{\partial D_{x0}}{\partial t} = \frac{\omega \varepsilon V_0}{d} \sin \omega t \tag{1.105}$$

$$\mathbf{H}_1 = \frac{\omega \varepsilon V_0 z}{d} \sin \omega t \, \mathbf{a}_y \tag{1.106}$$

where we have also satisfied the boundary condition at $z=0$ by choosing the constant of integration such that $\left[H_{y1}\right]_{z=0}$ is zero, and the subscript 1 denotes that the amplitude of the field is of the first power in ω. Note that the amplitude of H_{y1} varies linearly with z, from zero at $z=0$ to a maximum at $z=-l$.

We stop the solution here, because continuing the process by substituting Eq. (1.106) into Maxwell's curl equation for $\mathbf{E}$, Eq. (1.27) to obtain the resulting electric field will yield a term having amplitude proportional to the second power in ω. This simply means that the fields given as a pair by Eqs. (1.104) and (1.106) do not satisfy Eq. (1.27) and hence are not complete solutions to Maxwell's equations. The complete solutions are obtained by solving Maxwell's equations simultaneously and subject to the boundary conditions for the given problem.

Proceeding further, we obtain the current drawn from the voltage source to be

$$\begin{aligned} I_g(t) &= w\left[H_{y1}\right]_{z=-l} \\ &= -\omega\left(\frac{\varepsilon w l}{d}\right) V_0 \sin \omega t \\ &= C\frac{dV_g(t)}{dt} \end{aligned} \tag{1.107}$$

or,

$$\bar{I}_g = j\omega C \bar{V}_g \tag{1.108}$$

where $C=\varepsilon w l/d$ is the capacitance of the arrangement obtained from static field considerations. Thus, the input admittance of the structure is $j\omega C$ so that its low-frequency input behavior is essentially that of a single capacitor of value same as that found from static field analysis of the structure. Indeed, from considerations of power flow, using Poynting's theorem, we obtain the power flowing into the structure to be

$$\begin{aligned} P_{\text{in}} &= wd\left[E_{x0}H_{y1}\right]_{z=0} \\ &= -\left(\frac{\varepsilon w l}{d}\right)\omega V_0^2 \sin \omega t \cos \omega t \\ &= \frac{d}{dt}\left(\frac{1}{2}CV_g^2\right) \end{aligned} \tag{1.109}$$

© 2006 by Taylor & Francis Group, LLC

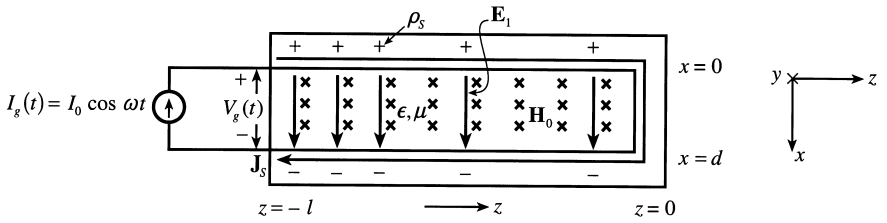


Figure 1.18 Magnetoquasistatic field analysis for the parallel-plate structure of Fig. 1.14.

which is consistent with the electric energy stored in the structure for the static case, as given by Eq. (1.79).

Magnetoquasistatic Fields

For magnetoquasistatic fields, we begin with the magnetic field having the spatial dependence of the static field solution for the given arrangement. An example is provided by the arrangement in Fig. 1.14a excited by a sinusoidally time-varying current source $I_g(t) = I_0 \cos \omega t$, instead of a direct current source, as shown by the cross-sectional view in Fig. 1.18. Then,

$$\mathbf{H}_0 = \frac{I_0}{w} \cos \omega t \, \mathbf{a}_y \tag{1.110}$$

where the subscript 0 again denotes that the amplitude of the field is of the zeroth power in ω. This results in an electric field in accordance with Maxwell's curl equation for $\mathbf{E}$, given by Eq. (1.27). Thus, we have for the geometry of the arrangement,

$$\frac{\partial E_{x1}}{\partial z} = -\frac{\partial B_{y0}}{\partial t} = \frac{\omega \mu I_0}{w} \sin \omega t \tag{1.111}$$

$$\mathbf{E}_1 = \frac{\omega \mu I_0 z}{w} \sin \omega t \, \mathbf{a}_x \tag{1.112}$$

where we have also satisfied the boundary condition at $z = 0$ by choosing the constant of integration such that $[E_{x1}]_{z=0} = 0$ is zero, and again the subscript 1 denotes that the amplitude of the field is of the first power in ω. Note that the amplitude of E_{x1} varies linearly with z, from zero at $z = 0$ to a maximum at $z = -l$.

As in the case of electroquasistatic fields, we stop the process here, because continuing it by substituting Eq. (1.112) into Maxwell's curl equation for $\mathbf{H}$, Eq. (1.28), to obtain the resulting magnetic field will yield a term having amplitude proportional to the second power in ω. This simply means that the fields given as a pair by Eqs. (1.110) and (1.112) do not satisfy Eq. (1.28), and hence are not complete solutions to Maxwell's equations. The complete solutions are obtained by solving Maxwell's equations simultaneously and subject to the boundary conditions for the given problem.

© 2006 by Taylor & Francis Group, LLC

Proceeding further, we obtain the voltage across the current source to be

$$\begin{aligned} V_g(t) &= d[E_{x1}]_{z=-l} \\ &= -\omega\left(\frac{\mu\, dl}{w}\right) I_0 \sin \omega t \\ &= L\frac{dI_g(t)}{dt} \end{aligned} \tag{1.113}$$

or

$$\bar{V}_g = j\omega L\, \bar{I}_g \tag{1.114}$$

where $L = \mu dl/w$ is the inductance of the arrangement obtained from static field considerations. Thus, the input impedance of the structure is $j\omega L$, such that its low-frequency input behavior is essentially that of a single inductor of value the same as that found from static field analysis of the structure. Indeed, from considerations of power flow, using Poynting's theorem, we obtain the power flowing into the structure to be

$$\begin{aligned} P_{\text{in}} &= wd\left[E_{x1} H_{y0}\right]_{z=-l} \\ &= -\left(\frac{\mu dl}{w}\right)\omega I_0^2 \sin \omega t \cos \omega t \\ &= \frac{d}{dt}\left(\frac{1}{2} L I_g^2\right) \end{aligned} \tag{1.115}$$

which is consistent with the magnetic energy stored in the structure for the static case, as given by Eq. (1.92).

Quasistatic Fields in a Conductor

If the dielectric slab in the arrangement of Fig. 1.17 is conductive, as shown in Fig. 1.19a, then both electric and magnetic fields exist in the static case because of the conduction current, as discussed under electromagnetostatic fields in Sec. 1.3.2. Furthermore, the electric field of amplitude proportional to the first power in ω contributes to the creation of magnetic field of amplitude proportional to the first power in ω, in addition to that from electric field of amplitude proportional to the zeroth power in ω.

Thus, using the results from the static field analysis for the arrangement of Fig. 1.15, we have for the arrangement of Fig. 1.19a

$$\mathbf{E}_0 = \frac{V_0}{d} \cos \omega t\, \mathbf{a}_x \tag{1.116}$$

$$\mathbf{J}_{c0} = \sigma \mathbf{E}_0 = \frac{\sigma V_0}{d} \cos \omega t\, \mathbf{a}_x \tag{1.117}$$

$$\mathbf{H}_0 = -\frac{\sigma V_0 z}{d} \cos \omega t\, \mathbf{a}_y \tag{1.118}$$

© 2006 by Taylor & Francis Group, LLC

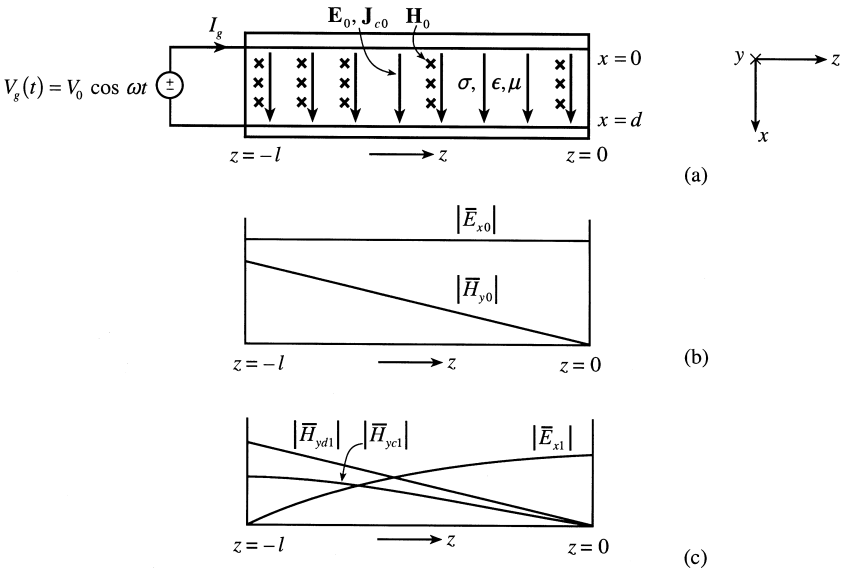


Figure 1.19 (a) Zero-order fields for the parallel-plate structure of Fig. 1.15. (b) Variations of amplitudes of the zero-order fields along the structure. (c) Variations of amplitudes of the first-order fields along the structure.

as depicted in the figure. Also, the variations with z of the amplitudes of E_{x0} and H_{y0} are shown in Fig. 1.19b.

The magnetic field given by Eq. (1.118) gives rise to an electric field having amplitude proportional to the first power in ω, in accordance with Maxwell's curl equation for **E**, Eq. (1.27). Thus

$$\frac{\partial E_{x1}}{\partial z} = -\frac{\partial B_{y0}}{\partial t} = -\frac{\omega\mu\sigma V_0 z}{d} \sin \omega t \tag{1.119}$$

$$E_{x1} = -\frac{\omega\mu\sigma V_0}{2d}\left(z^2 - l^2\right) \sin \omega t \tag{1.120}$$

where we have also made sure that the boundary condition at $z = -l$ is satisfied. This boundary condition requires that E_x be equal to V_g/d at $z = -l$. Since this is satisfied by E_{x0} alone, it follows that E_{x1} must be zero at $z = -l$.

The electric field given by Eq. (1.116) and that given by Eq. (1.120) together give rise to a magnetic field having terms with amplitudes proportional to the first power in ω, in accordance with Maxwell's curl equation for **H**, Eq. (1.28). Thus

$$\begin{aligned}\frac{\partial H_{y1}}{\partial z} &= -\sigma E_{x1} - \varepsilon \frac{\partial E_{x0}}{\partial t} \\ &= \frac{\omega\mu\sigma^2 V_0}{2d}\left(z^2 - l^2\right) \sin \omega t + \frac{\omega\varepsilon V_0}{d} \sin \omega t\end{aligned} \tag{1.121}$$

$$H_{y1} = \frac{\omega\mu\sigma^2 V_0\left(z^3 - 3zl^2\right)}{6d} \sin \omega t + \frac{\omega\varepsilon V_0 z}{d} \sin \omega t \tag{1.122}$$

© 2006 by Taylor & Francis Group, LLC

where we have also made sure that the boundary condition at $z=0$ is satisfied. This boundary condition requires that H_y be equal to zero at $z=0$, which means that all of its terms must be zero at $z=0$. Note that the first term on the right side of Eq. (1.122) is the contribution from the conduction current in the material resulting from E_{x1} and the second term is the contribution from the displacement current resulting from E_{x0}. Denoting these to be H_{yc1} *and* H_{yd1}, respectively, we show the variations with z of the amplitudes of all the field components having amplitudes proportional to the first power in ω, in Fig. 1.19c.

Now, adding up the contributions to each field, we obtain the total electric and magnetic fields up to and including the terms with amplitudes proportional to the first power in ω to be

$$E_x = \frac{V_0}{d}\cos\omega t - \frac{\omega\mu\sigma V_0}{2d}\left(z^2 - l^2\right)\sin\omega t \tag{1.123a}$$

$$H_y = -\frac{\sigma V_0 z}{d}\cos\omega t + \frac{\omega\varepsilon V_0 z}{d}\sin\omega t + \frac{\omega\mu\sigma^2 V_0\left(z^3 - 3zl^2\right)}{6d}\sin\omega t \tag{1.123b}$$

or

$$\bar{E}_x = \frac{\bar{V}_g}{d} + j\omega\frac{\mu\sigma}{2d}\left(z^2 - l^2\right)\bar{V}_g \tag{1.124a}$$

$$\bar{H}_y = -\frac{\sigma z}{d}\bar{V}_g - j\omega\frac{\varepsilon z}{d}\bar{V}_g - j\omega\frac{\mu\sigma^2\left(z^3 - 3zl^2\right)}{6d}\bar{V}_g \tag{1.124b}$$

Finally, the current drawn from the voltage source is given by

$$\begin{aligned}\bar{I}_g &= w\left[\bar{H}_y\right]_{z=-l} \\ &= \left(\frac{\sigma wl}{d} + j\omega\frac{\varepsilon wl}{d} - j\omega\frac{\mu\sigma^2 wl^3}{3d}\right)\bar{V}_g\end{aligned} \tag{1.125}$$

The input admittance of the structure is given by

$$\begin{aligned}\bar{Y}_{\text{in}} &= \frac{\bar{I}_g}{\bar{V}_g} = j\omega\frac{\varepsilon wl}{d} + \frac{\sigma wl}{d}\left(1 - j\omega\frac{\mu\sigma l^2}{3}\right) \\ &\approx j\omega\frac{\varepsilon wl}{d} + \frac{1}{(d/\sigma wl)[1 + j\omega(\mu\sigma l^2/3)]}\end{aligned} \tag{1.126}$$

where we have used the approximation $[1 + j\omega(\mu\sigma l^2/3)]^{-1} \approx [1 - j\omega(\mu\sigma l^2/3)]$. Proceeding further, we have

$$\begin{aligned}\bar{Y}_{\text{in}} &= j\omega\frac{\varepsilon wl}{d} + \frac{1}{(d/\sigma wl) + j\omega(\mu dl/3w)} \\ &= j\omega C + \frac{1}{R + j\omega L_i}\end{aligned} \tag{1.127}$$

© 2006 by Taylor & Francis Group, LLC

where $C = \varepsilon wl/d$ is the capacitance of the structure if the material is a perfect dielectric, $R = d/\sigma wl$ is the resistance of the structure, and $L_i = \mu dl/3w$ is the internal inductance of the structure, all computed from static field analysis of the structure.

The equivalent circuit corresponding to Eq. (1.127) consists of capacitance C in parallel with the series combination of resistance R and internal inductance L_i, same as in Fig. 1.16. Thus, the low-frequency input behavior of the structure is essentially the same as that of the equivalent circuit of Fig. 1.16, with the understanding that its input admittance must also be approximated to first-order terms. Note that for $\sigma = 0$, the input admittance of the structure is purely capacitive. For nonzero σ, a critical value of σ equal to $\sqrt{3\varepsilon/\mu l^2}$ exists for which the input admittance is purely conductive. For values of σ smaller than the critical value, the input admittance is complex and capacitive, and for values of σ larger than the critical value, the input admittance is complex and inductive.

1.3.4. Waves and the Distributed Circuit Concept

In Sec. 1.3.3, we have seen that quasistatic field analysis of a physical structure provides information concerning the low-frequency input behavior of the structure. As the frequency is increased beyond that for which the quasistatic approximation is valid, terms in the infinite series solutions for the fields beyond the first-order terms need to be included. While one can obtain equivalent circuits for frequencies beyond the range of validity of the quasistatic approximation by evaluating the higher order terms, no further insight is gained through that process, and it is more straightforward to obtain the exact solution by resorting to simultaneous solution of Maxwell's equations when a closed form solution is possible.

Wave Equation and Solutions

Let us, for simplicity, consider the structures of Figs. 1.17 and 1.18, for which the material between the plates is a perfect dielectric ($\sigma = 0$). Then, regardless of the termination at $z = 0$, the equations to be solved are

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\mu \frac{\partial \mathbf{H}}{\partial t} \tag{1.128a}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} = \varepsilon \frac{\partial \mathbf{E}}{\partial t} \tag{1.128b}$$

For the geometry of the arrangements, $\mathbf{E} = E_x(z, t)\mathbf{a}_x$ and $\mathbf{H} = H_y(z, t)\mathbf{a}_y$, so that Eqs. (1.128a) and (1.128b) simplify to

$$\frac{\partial E_x}{\partial z} = -\mu \frac{\partial H_y}{\partial t} \tag{1.129a}$$

$$\frac{\partial H_y}{\partial z} = -\varepsilon \frac{\partial E_x}{\partial t} \tag{1.129b}$$

© 2006 by Taylor & Francis Group, LLC

Combining the two equations by eliminating H_y, we obtain

$$\frac{\partial^2 E_x}{\partial z^2} = \mu\varepsilon \frac{\partial^2 E_x}{\partial t^2} \tag{1.130}$$

which is the *wave equation*. It has solutions of the form

$$E_x(z, t) = A \cos\omega\left(t - \sqrt{\mu\varepsilon}\, z + \phi^+\right) + B \cos\omega\left(t + \sqrt{\mu\varepsilon}\, z + \phi^-\right) \tag{1.131}$$

The terms on the right side correspond to traveling waves propagating in the $+z$ and $-z$ directions, which we shall call the (+) and (−) waves, respectively, with the velocity $1/\sqrt{\mu\varepsilon}$, or $c/\sqrt{\mu_r\varepsilon_r}$, where $c = 1/\sqrt{\mu_0\varepsilon_0}$ is the velocity of light in free space. This can be seen by setting the derivative of the argument of the cosine function in each term equal to zero or by plotting each term versus z for a few values of t, as shown in Fig. 1.20a and b for the (+) and (−) waves, respectively. The corresponding solution for H_y is given by

$$H_y(z, t) = \frac{1}{\sqrt{\mu/\varepsilon}}\left[A \cos\omega\left(t - \sqrt{\mu\varepsilon}\, z + \phi^+\right) - B \cos\omega\left(t + \sqrt{\mu\varepsilon}\, z + \phi^-\right)\right] \tag{1.132}$$

For sinusoidal waves, which is the case at present, the velocity of propagation is known as the *phase velocity*, denoted by v_p, since it is the velocity with which a constant phase surface moves in the direction of propagation. The quantity $\omega\sqrt{\mu\varepsilon}$ is the magnitude

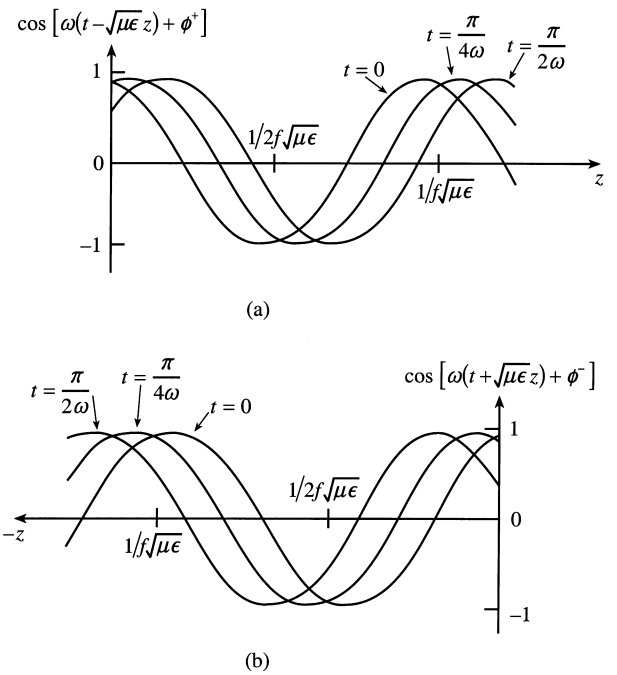


Figure 1.20 Plots of (a) $\cos\left[\omega\left(t - \sqrt{\mu\varepsilon}\, z\right) + \phi^+\right]$and (b) $\cos\left[\omega\left(t + \sqrt{\mu\varepsilon} z\right) + \phi^-\right]$, versus z, for a few values of t.

© 2006 by Taylor & Francis Group, LLC

of the rate of change of phase at a fixed time t, for either wave. It is known as the *phase constant* and is denoted by the symbol β. The quantity $\sqrt{\mu/\varepsilon}$, which is the ratio of the electric field intensity to the magnetic field intensity for the (+) wave, and the negative of such ratio for the (−) wave, is known as the *intrinsic impedance* of the medium. It is denoted by the symbol η. Thus, the phasor electric and magnetic fields can be written as

$$\bar{E}_x = \bar{A}e^{-j\beta z} + \bar{B}e^{j\beta z} \tag{1.133}$$

$$\bar{H}_y = \frac{1}{\eta}\left(\bar{A}e^{-j\beta z} - \bar{B}e^{j\beta z}\right) \tag{1.134}$$

We may now use the boundary conditions for a given problem and obtain the specific solution for that problem. For the arrangement of Fig. 1.17, the boundary conditions are $\bar{H}_y = 0$ at $z = 0$ and $\bar{E}_x = \bar{V}_g/d$ at $z = -l$. We thus obtain the particular solution for that arrangement to be

$$\bar{E}_x = \frac{\bar{V}_g}{d\ \cos\beta l}\cos\beta z \tag{1.135}$$

$$\bar{H}_y = \frac{-j\bar{V}_g}{\eta d\ \cos\beta l}\sin\beta z \tag{1.136}$$

which correspond to complete standing waves, resulting from the superposition of (+) and (−) waves of equal amplitude. Complete standing waves are characterized by pure half-sinusoidal variations for the amplitudes of the fields, as shown in Fig. 1.21. For values of z at which the electric field amplitude is a maximum, the magnetic field amplitude is zero, and for values of z at which the electric field amplitude is zero, the magnetic field amplitude is a maximum. The fields are also out of phase in time, such that at any value of z, the magnetic field and the electric field differ in phase by $t = \pi/2\omega$.

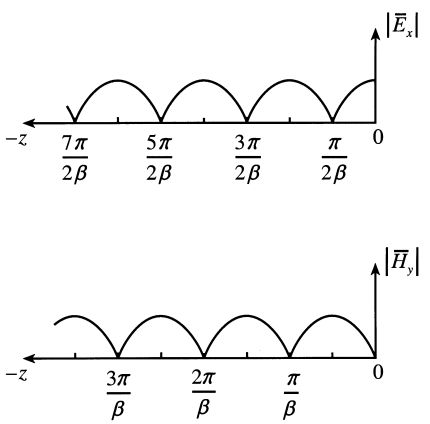


Figure 1.21 Standing wave patterns for the fields for the structure of Fig. 1.17.

© 2006 by Taylor & Francis Group, LLC

Now, the current drawn from the voltage source is given by

$$\begin{aligned}\bar{I}_g &= w\left[\bar{H}_y\right]_{z=-l} \\ &= \frac{jw\bar{V}_g}{\eta d}\tan\beta l\end{aligned} \tag{1.137}$$

so that the input impedance of the structure is

$$\bar{Y}_{\text{in}} = \frac{\bar{I}_g}{\bar{V}_g} = j\frac{w}{\eta d}\tan\beta l \tag{1.138}$$

which can be expressed as a power series in βl. In particular, for $\beta l < \pi/2$,

$$\bar{Y}_{\text{in}} = j\frac{w}{\eta d}\left[\beta l + \frac{(\beta l)^3}{3} + \frac{2(\beta l)^5}{15} + \cdots\right] \tag{1.139}$$

The first term on the right side can be identified as belonging to the quasistatic approximation. Indeed for $\beta l \ll 1$, the higher order terms can be neglected, and

$$\begin{aligned}\bar{Y}_{\text{in}} &\approx \frac{jw}{\eta d}(\beta l) \\ &= j\omega\left(\frac{\varepsilon w l}{d}\right)\end{aligned} \tag{1.140}$$

same as that given by Eq. (1.111).

It can now be seen that the condition $\beta l \ll 1$ dictates the range of validity for the quasistatic approximation for the input behavior of the structure. In terms of the frequency f of the source, this condition means that $f \ll v_p/2\pi l$, or in terms of the period $T = 1/f$, it means that $T \gg 2\pi(l/v_p)$. Thus, as already mentioned, quasistatic fields are low-frequency approximations of time-varying fields that are complete solutions to Maxwell's equations, which represent wave propagation phenomena and can be approximated to the quasistatic character only when the times of interest are much greater than the propagation time, l/v_p, corresponding to the length of the structure. In terms of space variations of the fields at a fixed time, the wavelength $\lambda\,(=2\pi/\beta)$, which is the distance between two consecutive points along the direction of propagation between which the phase difference is 2π, must be such that $l \ll \lambda/2\pi$; thus, the physical length of the structure must be a small fraction of the wavelength. In terms of amplitudes of the fields, what this means is that over the length of the structure, the field amplitudes are fractional portions of the first one-quarter sinusoidal variations at the $z = 0$ end in Fig. 1.21, with the boundary conditions at the two ends of the structure always satisfied. Thus, because of the cos βz dependence of $\bar{E}_x$ on z, the electric field amplitude is essentially a constant, whereas because of the sin βz dependence of $\bar{H}_y$ on z, the magnetic field amplitude varies linearly with z. These are exactly the nature of the variations of the zero-order electric field and the first-order magnetic field, as discussed under electroquasistatic fields in Sec. 1.3.3.

© 2006 by Taylor & Francis Group, LLC

For frequencies slightly beyond the range of validity of the quasistatic approximation, we can include the second term in the infinite series on the right side of Eq. (1.139) and deduce the equivalent circuit in the following manner.

$$\begin{aligned}\bar{Y}_{\text{in}} &\approx j\frac{w}{\eta d}\left[\beta l + \frac{(\beta l)^3}{3}\right] \\ &= j\omega\left(\frac{\varepsilon w l}{d}\right)\left[1 + \left(\omega\frac{\varepsilon w l}{d}\right)\left(\omega\frac{\mu d l}{3w}\right)\right]\end{aligned} \tag{1.141}$$

or

$$\begin{aligned}\bar{Z}_{\text{in}} &= \frac{1}{j\omega(\varepsilon w l/d)[1 + \omega(\varepsilon w l/d)(\omega\mu d l/3w)]} \\ &\approx \frac{1}{j\omega(\varepsilon w l/d)} + j\omega(\mu d l/3w)\end{aligned} \tag{1.142}$$

Thus the input behavior is equivalent to that of a capacitor of value same as that for the quasistatic approximation in series with an inductor of value 1/3 times the inductance found under the quasistatic approximation for the same arrangement but shorted at $z = 0$, by joining the two parallel plates. This series inductance is familiarly known as the *stray inductance*. But, all that has occurred is that the fractional portion of the sinusoidal variations of the field amplitudes over the length of the structure has increased, because the wavelength has decreased. As the frequency of the source is further increased, more and more terms in the infinite series need to be included, and the equivalent circuit becomes more and more involved. But throughout all this range of frequencies, the overall input behavior is still capacitive, until a frequency is reached when βl crosses the value $\pi/2$ and $\tan \beta l$ becomes negative, and the input behavior changes to inductive! In fact, a plot of $\tan \beta l$ versus f, shown in Fig. 1.22, indicates that as the frequency is varied, the input behavior alternates between capacitive and inductive, an observation unpredictable without the complete solutions to Maxwell's equations. At the frequencies at which the input behavior changes from capacitive to inductive, the input admittance becomes infinity (short-circuit condition). The field amplitude variations along the length of the structure are then exactly odd integer multiples of one-quarter sinusoids. At the frequencies at

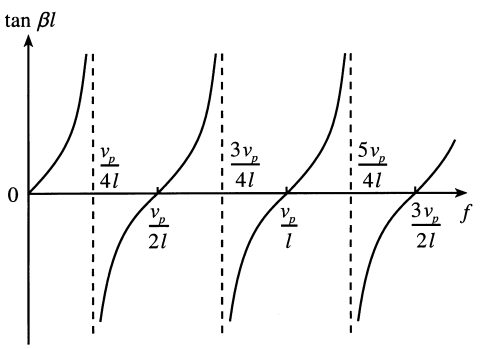


Figure 1.22 Frequency dependence of $\tan \beta l$.

© 2006 by Taylor & Francis Group, LLC

which the input behavior changes from inductive to capacitive, the input admittance becomes zero (open-circuit condition). The field amplitude variations along the length of the structure are then exactly even integer multiple of one-quarter sinusoids, or integer multiples of one-half sinusoids.

Turning now to the arrangement of Fig. 1.18, the boundary conditions are $\bar{E}_x = 0$ at $z = 0$ and $\bar{H}_y = \bar{I}_g / w$ at $z = -l$. We thus obtain the particular solution for that arrangement to be

$$\bar{E}_x = -\frac{j\eta \bar{I}_g}{w \cos \beta l} \sin \beta z \tag{1.143}$$

$$\bar{H}_y = \frac{\bar{I}_g}{w \cos \beta l} \cos \beta z \tag{1.144}$$

which, once again, correspond to complete standing waves, resulting from the superposition of (+) and (−) waves of equal amplitude, and characterized by pure half-sinusoidal variations for the amplitudes of the fields, as shown in Fig. 1.23, which are of the same nature as in Fig. 1.21, except that the electric and magnetic fields are interchanged.

Now, the voltage across the current source is given by

$$\begin{aligned} \bar{V}_g &= d\left[\bar{E}_x\right]_{z=-l} \\ &= \frac{j\eta d \bar{I}_g}{w} \tan \beta l \end{aligned} \tag{1.145}$$

so that the input impedance of the structure is

$$\bar{Z}_{\text{in}} = \frac{\bar{V}_g}{\bar{I}_g} = j\frac{\eta d}{w} \tan \beta l \tag{1.146}$$

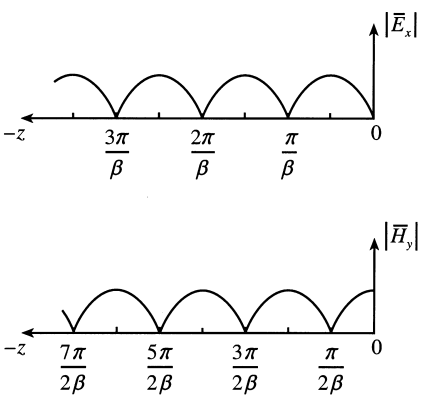


Figure 1.23 Standing wave patterns for the fields for the structure of Fig. 1.18.

© 2006 by Taylor & Francis Group, LLC

which can be expressed as a power series in βl. In particular, for $\beta l < \pi/2$,

$$\bar{Z}_{in} = j\frac{\eta d}{w}\left[\beta l + \frac{(\beta l)^3}{3} + \frac{2(\beta l)^5}{15} + \cdots\right] \tag{1.147}$$

Once again, the first term on the right side can be identified as belonging to the quasistatic approximation. Indeed for $\beta l \ll 1$,

$$\begin{aligned}\bar{Z}_{in} &\approx \frac{\eta d}{w}(\beta l) \\ &= j\omega\left(\frac{\mu d l}{w}\right)\end{aligned} \tag{1.148}$$

same as that given by Eq. (1.118), and all the discussion pertinent to the condition for the validity of the quasistatic approximation for the structure of Fig. 1.17 applies also to the structure of Fig. 1.18, with the roles of the electric and magnetic fields interchanged. For $l \ll \lambda/2\pi$, the field amplitudes over the length of the structure are fractional portions of the first one-quarter sinusoidal variations at the $z = 0$ end in Fig. 1.23, with the boundary conditions at the two ends always satisfied. Thus, because of the $\cos \beta z$ dependence of $\bar{H}_y$ on z, the magnetic field amplitude is essentially a constant, whereas because of the $\sin \beta z$ dependence of $\bar{E}_x$ on z, the electric field amplitude varies linearly with z. These are exactly the nature of the variations of the zero-order magnetic field and the first-order electric field, as discussed under magnetoquasistatic fields in Sec. 1.3.3.

For frequencies slightly beyond the range of validity of the quasistatic approximation, we can include the second term in the infinite series on the right side of Eq. (1.147) and deduce the equivalent circuit in the following manner.

$$\begin{aligned}\bar{Z}_{\text{in}} &\approx \frac{j\eta d}{w}\left[\beta l + \frac{(\beta l)^3}{3}\right] \\ &= j\omega\left(\frac{\mu d l}{w}\right)\left[1 + \left(\omega\frac{\mu d l}{w}\right)\left(\omega\frac{\varepsilon w l}{3d}\right)\right]\end{aligned} \tag{1.149}$$

or

$$\begin{aligned}\bar{Y}_{\text{in}} &= \frac{1}{j\omega(\mu d l/w)[1 + (\omega\mu d l/w)(\omega\varepsilon w l/3d)]} \\ &\approx \frac{1}{j\omega(\mu d l/w)} + j\omega\left(\frac{\varepsilon w l}{3d}\right)\end{aligned} \tag{1.150}$$

Thus the input behavior is equivalent to that of an inductor of value same as that for the quasistatic approximation in parallel with a capacitor of value 1/3 times the capacitance found under the quasistatic approximation for the same arrangement but open at $z = 0$, without the two plates joined. This parallel capacitance is familiarly known as the *stray capacitance*. But again, all that has occurred is that the fractional portion of the sinusoidal variations of the field amplitudes over the length of the structure has increased, because the wavelength has decreased. As the frequency of the source is further increased, more and more terms in the infinite series need to be included and the equivalent circuit becomes

© 2006 by Taylor & Francis Group, LLC

more and more involved. But throughout all this range of frequencies, the overall input behavior is still inductive, until a frequency is reached when βl crosses the value $\pi/2$ and $\tan \beta l$ becomes negative and the input behavior changes to capacitive. In fact, the plot of $\tan \beta l$ versus f, shown in Fig. 1.22, indicates that as the frequency is varied, the input behavior alternates between inductive and capacitive, an observation unpredictable without the complete solutions to Maxwell's equations. At the frequencies at which the input behavior changes from inductive to capacitive, the input impedance becomes infinity (open-circuit condition). The field amplitude variations along the length of the structure are then exactly odd integer multiples of one-quarter sinusoids. At the frequencies at which the input behavior changes from capacitive to inductive, the input impedance becomes zero (short-circuit condition). The field amplitude variations along the length of the structure are then exactly even integer multiples of one-quarter sinusoids, or integer multiples of one-half sinusoids.

Distributed Circuit Concept

We have seen that, from the circuit point of view, the structure of Fig. 1.13 behaves like a capacitor for the static case and the capacitive character is essentially retained for its input behavior for sinusoidally time-varying excitation at frequencies low enough to be within the range of validity of the quasistatic approximation. Likewise, we have seen that from a circuit point of view, the structure of Fig. 1.14 behaves like an inductor for the static case and the inductive character is essentially retained for its input behavior for sinusoidally time-varying excitation at frequencies low enough to be within the range of validity of the quasistatic approximation. For both structures, at an arbitrarily high enough frequency, the input behavior can be obtained only by obtaining complete (wave) solutions to Maxwell's equations, subject to the appropriate boundary conditions. The question to ask then is whether there is a circuit equivalent for the structure itself, independent of the termination, that is representative of the phenomenon taking place along the structure and valid at any arbitrary frequency, to the extent that the material parameters themselves are independent of frequency? The answer is, yes, under certain conditions, giving rise to the concept of the *distributed circuit*.

To develop and discuss the concept of the distributed circuit using a more general case than that allowed by the arrangements of Figs. 1.13 and 1.14, let us consider the case of the structure of Fig. 1.15 driven by a sinusoidally time-varying source, as in Fig. 1.19a. Then the equations to be solved are

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\mu \frac{\partial \mathbf{H}}{\partial t} \tag{1.151a}$$

$$\nabla \times \mathbf{H} = \mathbf{J}_c + \frac{\partial \mathbf{D}}{\partial t} = \sigma \mathbf{E} + \varepsilon \frac{\partial \mathbf{E}}{\partial t} \tag{1.151b}$$

For the geometry of the arrangement, $\mathbf{E} = E_x(z, t)\mathbf{a}_x$ and $\mathbf{H} = H_y(z, t)\mathbf{a}_y$, so that Eqs. (1.151a) and (1.151b) simplify to

$$\frac{\partial E_x}{\partial z} = -\mu \frac{\partial H_y}{\partial t} \tag{1.152a}$$

$$\frac{\partial H_y}{\partial z} = -\sigma E_x - \varepsilon \frac{\partial E_x}{\partial t} \tag{1.152b}$$

© 2006 by Taylor & Francis Group, LLC

Now, since E_z and H_z are zero, we can, in a given $z =$ constant plane, uniquely define a voltage between the plates in terms of the electric field intensity in that plane and a current crossing that plane in one direction on the top plate and in the opposite direction on the bottom plate in terms of the magnetic field intensity in that plane. These are given by

$$V(z, t) = dE_x(z, t) \tag{1.153a}$$

$$I(z, t) = wH_y(z, t) \tag{1.153b}$$

Substituting Eqs. (1.153a) and (1.153b) in Eqs. (1.152a) and (1.152b), and rearranging, we obtain

$$\frac{\partial V(z, t)}{\partial z} = -\left[\frac{\mu d}{w}\right]\frac{\partial I(z, t)}{\partial t} \tag{1.154a}$$

$$\frac{\partial I(z, t)}{\partial z} = -\left[\frac{\sigma w}{d}\right]V(z, t) - \left[\frac{\varepsilon w}{d}\right]\frac{\partial V(z, t)}{\partial t} \tag{1.154b}$$

Writing the derivates with respect to z on the left sides of the equations in terms of limits as $\Delta z \to 0$, and multiplying by Δz on both sides of the equations provides the equivalent circuit for a section of length Δz of the structure, as shown in Fig. 1.24, in which the quantities $\mathcal{L}$, $\mathcal{C}$, and $\mathcal{G}$, given by

$$\mathcal{L} = \frac{\mu d}{w} \tag{1.155a}$$

$$\mathcal{C} = \frac{\varepsilon w}{d} \tag{1.155b}$$

$$\mathcal{G} = \frac{\sigma w}{d} \tag{1.155c}$$

are the inductance per unit length, capacitance per unit length, and conductance per unit length, respectively, of the structure, all computed from static field analysis, except that now they are expressed in terms of "per unit length" and not for the entire structure in a "lump." It then follows that the circuit representation of the entire structure consists of an infinite number of such sections in cascade, as shown in Fig. 1.25. Such a circuit is known as a *distributed circuit*. The distributed circuit notion arises from the fact that the

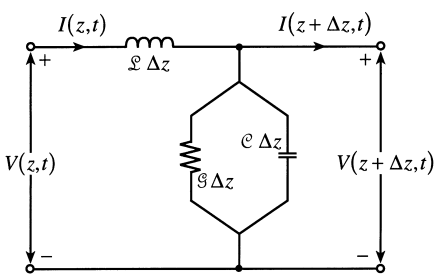


Figure 1.24 Circuit equivalent for Eqs. (1.159a and b), in the limit $\Delta z \to 0$.

© 2006 by Taylor & Francis Group, LLC

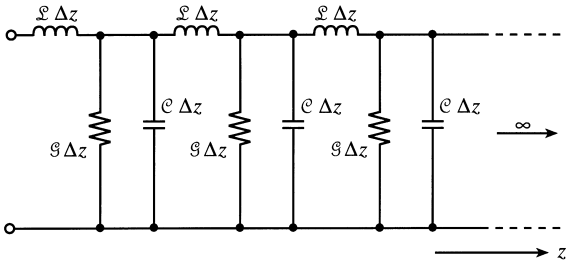


Figure 1.25 Distributed circuit representation of the structure of Fig. 1.19a.

inductance, capacitance, and conductance are distributed uniformly and overlappingly along the structure. A physical interpretation of the distributed-circuit concept follows from energy considerations, based on the properties that inductance, capacitance, and conductance are elements associated with energy storage in the magnetic field, energy storage in the electric field, and power dissipation due to conduction current flow, in the material. Since these phenomena occur continuously and overlappingly along the structure, the inductance, capacitance, and conductance must be distributed uniformly and overlappingly along the structure.

A physical structure for which the distributed circuit concept is applicable is familiarly known as a *transmission line*. The parallel-plate arrangement of Figs. 1.13–1.15 is a special case of a transmission line, known as the *parallel-plate* line, in which the waves are called *uniform plane waves*, since the fields are uniform in the $z =$ constant planes. In general, a transmission line consists of two parallel conductors having arbitrary cross sections and the waves are transverse electromagnetic, or TEM, waves, for which the fields are nonuniform in the $z =$ constant planes but satisfying the property of both electric and magnetic fields having no components along the direction of propagation, that is, parallel to the conductors. For waves for which the electric field has a component along the direction of propagation but the magnetic field does not, as is the case for transverse magnetic or TM waves, the current on the conductors crossing a given transverse plane cannot be expressed uniquely in terms of the magnetic field components in that plane. Likewise, for waves for which the magnetic field has a component along the direction of propagation but the electric field does not, as is the case for transverse electric or TE waves, the voltage between the conductors in a given transverse plane cannot be expressed uniquely in terms of the electric field components in that plane. Structures which support TM and TE waves are generally known as *waveguides*, although transmission lines are also waveguides in the sense that TEM waves are guided parallel to the conductors of the line.

All transmission lines having perfect conductors are governed by the equations

$$\frac{\partial V(z,\,t)}{\partial z} = -\mathcal{L}\,\frac{\partial I(z,\,t)}{\partial t} \tag{1.156a}$$

$$\frac{\partial I(z,\,t)}{\partial z} = -\mathcal{G}V(z,\,t) - \mathcal{C}\frac{\partial V(z,\,t)}{\partial t} \tag{1.156b}$$

which are known as the *transmission-line equations*. The values of $\mathcal{L}$, $\mathcal{C}$, and $\mathcal{G}$ differ from one line to another, and depend on the cross-sectional geometry of the conductors. For the

© 2006 by Taylor & Francis Group, LLC

parallel-plate line, $\mathcal{L}$, $\mathcal{C}$, and $\mathcal{G}$ are given by Eqs. (1.155a), (1.155b), and (1.155c), respectively. Note that

$$\mathcal{L}\mathcal{C} = \mu\varepsilon \tag{1.157a}$$

$$\frac{\mathcal{G}}{\mathcal{C}} = \frac{\sigma}{\varepsilon} \tag{1.157b}$$

a set of relations, which is applicable to any line governed by Eqs. (1.156a) and (1.156b). Thus for a given set of material parameters, only one of the three parameters, $\mathcal{L}$, $\mathcal{C}$, and $\mathcal{G}$, is independent.

In practice, the conductors are imperfect, adding a resistance per unit length and additional inductance per unit length in the series branches of the distributed circuit. Although the waves are then no longer exactly TEM waves, the distributed circuit is commonly used for transmission lines with imperfect conductors. Another consideration that arises in practice is that the material parameters and hence the line parameters can be functions of frequency.

1.3.5. Hertzian Dipole Fields via the Thread of Statics–Quasistatics–Waves

In the preceding three sections, we have seen the development of solutions to Maxwell's equations, beginning with static fields and spanning the frequency domain from quasistatic approximations at low frequencies to waves for beyond quasistatics. In this section, we shall develop the solution for the electromagnetic field due to a Hertzian dipole by making use of the thread of statics–quasistatics–waves, as compared to the commonly used approach based on the magnetic vector potential, for a culminating experience of revisiting the fundamentals of engineering electromagnetics.

The Hertzian dipole is an elemental antenna consisting of an infinitesimally long piece of wire carrying an alternating current $I(t)$, as shown in Fig. 1.26. To maintain the current flow in the wire, we postulate two point charges $Q_1(t)$ and $Q_2(t)$ terminating the wire at its two ends, so that the law of conservation of charge is satisfied. Thus, if

$$I(t) = I_0 \cos \omega t \tag{1.158}$$

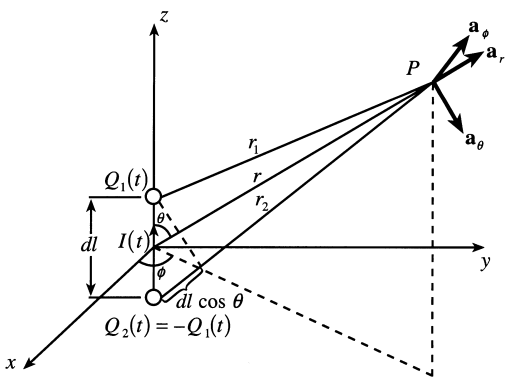


Figure 1.26 For the determination of the electromagnetic field due to the Hertzian dipole.

© 2006 by Taylor & Francis Group, LLC

then

$$Q_1(t) = \frac{I_0}{\omega} \sin \omega t \tag{1.159a}$$

$$Q_2(t) = -\frac{I_0}{\omega} \sin \omega t = -Q_1(t) \tag{1.159b}$$

For $d/dt = 0$, the charges are static and the current is zero. The field is simply the electrostatic field due to the electric dipole made up of $Q_1 = -Q_2 = Q_0$. Applying Eq. (1.70) to the geometry in Fig. 1.26, we write the electrostatic potential at the point P due to the dipole located at the origin to be

$$\Phi = \frac{Q_0}{4\pi\varepsilon}\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{1.160}$$

In the limit $dl \to 0$, keeping the dipole moment $Q_0(dl)$ fixed, we get

$$\Phi = \frac{Q_0(dl)\ \cos\theta}{4\pi\varepsilon r^2} \tag{1.161}$$

so that the electrostatic field at the point P due to the dipole is given by

$$\mathbf{E} = -\nabla\Phi = \frac{Q_0(dl)}{4\pi\varepsilon r^3}(2\ \cos\theta\ \mathbf{a}_r + \sin\theta\ \mathbf{a}_\theta) \tag{1.162}$$

With time variations in the manner $Q_1(t) = -Q_2(t) = Q_0 \sin \omega t$, so that $I_0 = \omega Q_0$, and at low frequencies, the situation changes to electroquasistatic with the electric field of amplitude proportional to the zeroth power in ω given by

$$\mathbf{E}_0 = \frac{Q_0(dl)\ \sin\omega t}{4\pi\varepsilon r^3}(2\ \cos\theta\ \mathbf{a}_r + \sin\theta\ \mathbf{a}_\theta) \tag{1.163}$$

The corresponding magnetic field of amplitude proportional to the first power in ω is given by the solution of

$$\nabla \times \mathbf{H}_1 = \frac{\partial \mathbf{D}_0}{\partial t} = \varepsilon \frac{\partial \mathbf{E}_0}{\partial t} \tag{1.164}$$

For the geometry associated with the arrangement, this reduces to

$$\begin{vmatrix} \dfrac{\mathbf{a}_r}{r^2\ \sin\theta} & \dfrac{\mathbf{a}_\theta}{r\ \sin\theta} & \dfrac{\mathbf{a}_\phi}{r} \\ \dfrac{\partial}{\partial r} & \dfrac{\partial}{\partial \theta} & 0 \\ 0 & 0 & r\ \sin\theta\ H_{\phi 1} \end{vmatrix} = \varepsilon \frac{\partial \mathbf{E}_0}{\partial t} \tag{1.165}$$

© 2006 by Taylor & Francis Group, LLC

so that

$$\mathbf{H}_1 = \frac{\omega Q_0(dl)\ \cos\omega t}{4\pi r^2}\ \sin\theta\ \mathbf{a}_\phi \tag{1.166}$$

To extend the solutions for the fields for frequencies beyond the range of validity of the quasistatic approximation, we recognize that the situation then corresponds to wave propagation. With the dipole at the origin, the waves propagate radially away from it so that the time functions sin ωt and cos ωt in Eqs. (1.163) and (1.166) need to be replaced by sin $(\omega t - \beta r)$ and cos $(\omega t - \beta r)$, respectively, where $\beta = \omega\sqrt{\mu\varepsilon}$ is the phase constant. Therefore, let us on this basis alone and without any other considerations, write the field expressions as

$$\mathbf{E} = \frac{I_0(dl)\ \sin(\omega t - \beta r)}{4\pi\varepsilon\omega r^3}(2\ \cos\theta\ \mathbf{a}_r + \sin\theta\ \mathbf{a}_\theta) \tag{1.167}$$

$$\mathbf{H} = \frac{I_0(dl)\ \cos(\omega t - \beta r)}{4\pi r^2}\ \sin\theta\ \mathbf{a}_\phi \tag{1.168}$$

where we have also replaced Q_0 by I_0/ω, and pose the question as to whether or not these expressions represent the solution for the electromagnetic field due to the Hertzian dipole. The answer is "no," since they do not satisfy Maxwell's curl equations

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\mu\frac{\partial \mathbf{H}}{\partial t} \tag{1.169a}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} = \varepsilon\frac{\partial \mathbf{E}}{\partial t} \tag{1.169b}$$

which can be verified by substituting them into the equations.

There is more than one way of resolving this discrepancy, but we shall here do it from physical considerations. Even a cursory look at the solutions for the fields given by Eqs. (1.167) and (1.168) points to the problem, since the Poynting vector $\mathbf{E} \times \mathbf{H}$ corresponding to them is proportional to $1/r^5$, and there is no real power flow associated with them because they are out of phase in ωt by $\pi/2$. But, we should expect that the fields contain terms proportional to $1/r$, which are in phase, from considerations of real power flow in the radial direction and from the behavior of the waves viewed locally over plane areas normal to the radial lines emanating from the Hertzian dipole, and electrically far from it ($\beta r \gg 1$), to be approximately that of uniform plane waves with the planes as their constant phase surfaces, as shown in Fig. 1.27.

To elaborate upon this, let us consider two spherical surfaces of radii r_a and r_b and centered at the dipole and insert a cone through these two surfaces such that its vertex is at the antenna, as shown in the Fig. 1.27. Then the power crossing any portion of the spherical surface of radius r_b must be the same as the power crossing the spherical surface of radius r_a inside the cone. Since these surface areas are proportional to the square of the radius and since the surface integral of the Poynting vector gives the power, the Poynting vector must have an r component proportional to $1/r^2$, and it follows that the solutions for E_θ and H_ϕ must contain terms proportional to $1/r$ and in phase.

© 2006 by Taylor & Francis Group, LLC

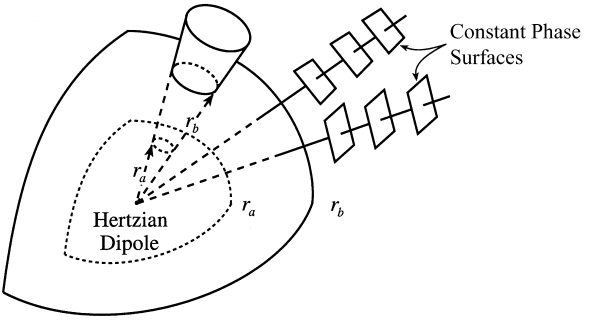


Figure 1.27 Radiation of electromagnetic waves far from the Hertzian dipole.

Thus let us modify the expression for **H** given by Eq. (1.168) by adding a second term containing $1/r$ in the manner

$$\mathbf{H} = \frac{I_0(dl)\ \sin\theta}{4\pi}\left[\frac{\cos(\omega t - \beta r)}{r^2} + \frac{A\ \cos(\omega t - \beta r + \delta)}{r}\right]\mathbf{a}_\phi \tag{1.170}$$

where A and δ are constants to be determined. Then, from Maxwell's curl equation for **H**, given by Eq. (1.169b), we obtain

$$\begin{aligned}\mathbf{E} = {} & \frac{2I_0(dl)\ \cos\theta}{4\pi\varepsilon\omega}\left[\frac{\sin(\omega t - \beta r)}{r^3} + \frac{A\ \sin(\omega t - \beta r + \delta)}{r^2}\right]\mathbf{a}_r \\ & + \frac{I_0(dl)\ \sin\ \theta}{4\pi\varepsilon\omega}\left[\frac{\sin(\omega t - \beta r)}{r^3} + \frac{\beta\ \sin(\omega t - \beta r)}{r^2}\right. \\ & \left. + \frac{A\beta\ \cos(\omega t - \beta r + \delta)}{r}\right]\mathbf{a}_\theta\end{aligned} \tag{1.171}$$

Now, substituting this in Maxwell's curl equation for **E** given by Eq. (1.169a), we get

$$\begin{aligned}\mathbf{H} = {} & \frac{I_0(dl)\ \sin\theta}{4\pi}\left[\frac{2\ \sin(\omega t - \beta r)}{\beta r^3} + \frac{2A\ \cos(\omega t - \beta r + \delta)}{\beta^2 r^3}\right. \\ & \left. + \frac{\cos(\omega t - \beta r)}{r^2} + \frac{A\ \cos(\omega t - \beta r + \delta)}{r}\right]\mathbf{a}_\phi\end{aligned} \tag{1.172}$$

But Eq. (1.172) must be the same as Eq. (1.170). Therefore, we set

$$\frac{2\ \sin(\omega t - \beta r)}{\beta r^3} + \frac{2A\ \cos(\omega t - \beta r + \delta)}{\beta^2 r^3} = 0 \tag{1.173}$$

© 2006 by Taylor & Francis Group, LLC

which gives us

$$\delta = \frac{\pi}{2} \tag{1.174}$$

$$A = \beta \tag{1.175}$$

Substituting Eqs. (1.174) and (1.175) in Eqs. (1.171) and (1.172), we then have the complete electromagnetic field due to the Hertzian dipole given by

$$\mathbf{E} = \frac{2I_0(dl)\ \cos\theta}{4\pi\varepsilon\omega}\left[\frac{\sin(\omega t-\beta r)}{r^3}+\frac{\beta\ \cos(\omega t-\beta r)}{r^2}\right]\mathbf{a}_r + \frac{I_0(dl)\ \sin\theta}{4\pi\varepsilon\omega}\left[\frac{\sin(\omega t-\beta r)}{r^3}+\frac{\beta\ \cos(\omega t-\beta r)}{r^2} - \frac{\beta^2\ \sin(\omega t-\beta r)}{r}\right]\mathbf{a}_\theta \tag{1.176}$$

$$\mathbf{H} = \frac{I_0(dl)\ \sin\theta}{4\pi}\left[\frac{\cos(\omega t-\beta r)}{r^2}-\frac{\beta\ \sin(\omega t-\beta r)}{r}\right]\mathbf{a}_\phi \tag{1.177}$$

Expressed in phasor form and with some rearrangement, the field components are given by

$$\bar{E}_r = \frac{2\beta^2\eta I_0(dl)\ \cos\theta}{4\pi}\left[-j\frac{1}{(\beta r)^3}+\frac{1}{(\beta r)^2}\right]e^{-j\beta r} \tag{1.178}$$

$$\bar{E}_\theta = \frac{\beta^2\eta I_0(dl)\ \sin\theta}{4\pi}\left[-j\frac{1}{(\beta r)^3}+\frac{1}{(\beta r)^2}+j\frac{1}{\beta r}\right]e^{-j\beta r} \tag{1.179}$$

$$\bar{H}_\phi = \frac{\beta^2 I_0(dl)\ \sin\theta}{4\pi}\left[\frac{1}{(\beta r)^2}+j\frac{1}{\beta r}\right]e^{-j\beta r} \tag{1.180}$$

The following observations are pertinent to these field expressions:

1. They satisfy all Maxwell's equations exactly.
2. For any value of r, the time-average value of the θ component of the Poynting vector is zero, and the time-average value of the r component of the Poynting vector is completely from the $1/r$ terms, thereby resulting in the time-average power crossing all possible spherical surfaces centered at the dipole to be the same.
3. At low frequencies such that $\beta r \ll 1$, the $1/(\beta r)^3$ terms dominate the $1/(\beta r)^2$ terms, which in turn dominate the $1/(\beta r)$ terms, and $e^{-j\beta r} \approx (1 - j\beta r)$, thereby reducing the field expressions to the phasor forms of the quasistatic approximations given by Eqs. (1.163) and (1.166).

Finally, they are the familiar expressions obtained by using the magnetic vector potential approach.

© 2006 by Taylor & Francis Group, LLC

REFERENCES

There is a multitude of textbooks on engineering electromagnetics, let alone electromagnetics, and it is difficult to prepare a list without inadvertently omitting some of them. Therefore, I have not attempted to include a bibliography of these books; instead, I refer the reader to his or her favorite book(s), while a student or later during the individual's career, and I list below my own books, which are referenced on the first page of this chapter.

1. Rao, N. N. *Basic Electromagnetics with Applications*; Prentice Hall: Englewood Cliffs, NJ, 1972.
2. Rao, N. N. *Elements of Engineering Electromagnetics*; Prentice Hall: Englewood Cliffs, NJ, 1977.
3. Rao, N. N. *Elements of Engineering Electromagnetics*; 2nd Ed.; Prentice Hall: Englewood Cliffs, NJ, 1987.
4. Rao, N. N. *Elements of Engineering Electromagnetics*; 3rd Ed.; Prentice Hall: Englewood Cliffs, NJ, 1991.
5. Rao, N. N. *Elements of Engineering Electromagnetics*; 4th Ed.; Prentice Hall; Englewood Cliffs, NJ, 1994.
6. Rao, N. N. *Elements of Engineering Electromagnetics*; 5th Ed.; Prentice Hall; Upper Saddle River, NJ, 2000.
7. Rao, N. N. *Elements of Engineering Electromagnetics*; 6th Ed.; Pearson Prentice Hall; Upper Saddle River, NJ, 2004.

© 2006 by Taylor & Francis Group, LLC

2
Applied Electrostatics

Mark N. Horenstein
Boston University
Boston, Massachusetts

2.1. INTRODUCTION

The term *electrostatics* brings visions of Benjamin Franklin, the "kite and key" experiment, Leyden jars, cat fur, and glass rods. These and similar experiments heralded the discovery of electromagnetism and were among some of the first recorded in the industrial age. The forces attributable to electrostatic charge have been known since the time of the ancient Greeks, yet the discipline continues to be the focus of much research and development. Most electrostatic processes fall into one of two categories. Sometimes, electrostatic charge produces a desired outcome, such as motion, adhesion, or energy dissipation. Electrostatic forces enable such diverse processes as laser printing, electrophotography, electrostatic paint spraying, powder coating, environmentally friendly pesticide application, drug delivery, food production, and electrostatic precipitation. Electrostatics is critical to the operation of micro-electromechanical systems (MEMS), including numerous microsensors, transducers, accelerometers, and the microfluidic "lab on a chip." These microdevices have opened up new vistas of discovery and have changed the way electronic circuits interface with the mechanical world. Electrostatic forces on a molecular scale lie at the core of nanodevices, and the inner workings of a cell's nucleus are also governed by electrostatics. A myriad of self-assembling nanodevices involving coulombic attraction and repulsion comprise yet another technology in which electrostatics plays an important role.

Despite its many useful applications, electrostatic charge is often a nuisance to be avoided. For example, sparks of electrostatic origin trigger countless accidental explosions every year and lead to loss of life and property. Less dramatically, static sparks can damage manufactured products such as electronic circuits, photographic film, and thin-coated materials. The transient voltage and current of a single spark event, called an *electrostatic discharge* (ESD), can render a semiconductor chip useless. Indeed, a billion-dollar industry specializing in the prevention or neutralization of ESD-producing electrostatic charge has of necessity evolved within the semiconductor industry to help mitigate this problem.

Unwanted electrostatic charge can also affect the production of textiles or plastics. Sheets of these materials, called *webs*, are produced on rollers at high speed. Electrostatic

© 2006 by Taylor & Francis Group, LLC

charge can cause webs to cling to rollers and jam production lines. Similarly, the sparks that result from accumulated charge can damage the product itself, either by exposing light-sensitive surfaces or by puncturing the body of the web.

This chapter presents the fundamentals that one needs in order to understand electrostatics as both friend and foe. We first define the electrostatic regime in the broad context of Maxwell's equations and review several fundamental concepts, including Coulomb's law, force-energy relations, triboelectrification, induction charging, particle electrification, and dielectric breakdown. We then examine several applications of electrostatics in science and industry and discuss some of the methods used to moderate the effects of unwanted charge.

2.2. THE ELECTROQUASISTATIC REGIME

Like all of electromagnetics, electrostatics is governed by Maxwell's equations, the elegant mathematical statements that form the basis for all that is covered in this book. True electrostatic systems are those in which all time derivatives in Maxwell's equations are exactly zero and in which forces of magnetic origin are absent. This limiting definition excludes numerous practical electrostatic-based applications. Fortunately, it can be relaxed while still capturing the salient features of the electrostatic domain. The *electroquasistatic* regime thus refers to those cases of Maxwell's equations in which fields and charge magnitudes may vary with time but in which the forces due to the electric field always dominate over the forces due to the magnetic field. At any given moment in time, an electroquasistatic field is identical to the field that would be produced were the relevant charges fixed at their instantaneous values and locations.

In order for a system to be electroquasistatic, two conditions must be true: First, any currents that flow within the system must be so small that the magnetic fields they produce generate negligible forces compared to coulombic forces. Second, any time variations in the electric field (or the charges that produce them) must occur so slowly that the effects of any induced magnetic fields are negligible. In this limit, the curl of E approaches zero, and the cross-coupling between E and H that would otherwise give rise to propagating waves is negligible. Thus one manifestation of the electroquasistatic regime is that the sources of the electric field produce no propagating waves.

The conditions for satisfying the electroquasistatic limit also can be quantified via dimensional analysis. The curl operator $\nabla\times$ has the dimensions of a reciprocal distance ΔL, while each time derivative dt in Maxwell's equations has the dimensions of a time Δt. Thus, considering Faraday's law:

$$\nabla \times \mathbf{E} = \frac{-\partial \mu \mathbf{H}}{\partial t} \tag{2.1}$$

the condition that the left-hand side be much greater than the right-hand side becomes dimensionally equivalent to

$$\frac{E}{\Delta L} \gg \frac{\mu H}{\Delta t} \tag{2.2}$$

© 2006 by Taylor & Francis Group, LLC

This same dimensional argument can be applied to Ampere's law:

$$\nabla \times \mathbf{H} = \frac{\partial \varepsilon \mathbf{E}}{\partial t} + \mathbf{J} \tag{2.3}$$

which, with $\mathbf{J} = 0$, leads to

$$\frac{H}{\Delta L} \gg \frac{\varepsilon E}{\Delta t} \tag{2.4}$$

Equation (2.4) for $\mathbf{H}$ can be substituted into Eq. (2.2), yielding

$$\frac{\mathbf{E}}{\Delta L} \gg \frac{\mu}{\Delta t} \frac{\varepsilon \mathbf{E} \Delta L}{\Delta t} \tag{2.5}$$

This last equation results in the dimensional condition that

$$\Delta L \ll \frac{\Delta t}{\sqrt{\mu \varepsilon}} \tag{2.6}$$

The quantity $1/\sqrt{\mu\varepsilon}$ is the propagation velocity of electromagnetic waves in the medium (i.e., the speed of light), hence $\Delta t/\sqrt{\mu\varepsilon}$ is the distance that a wave would travel after propagating for time Δt. If we interpret Δt as the period T of a possible propagating wave, then according to Eq. (2.6), the quasistatic limit applies if the length scale ΔL of the system is much smaller than the propagation wavelength at the frequency of excitation.

In the true electrostatic limit, the time derivatives are exactly zero, and Faraday's law Eq. (2.1) becomes

$$\nabla \times \mathbf{E} = 0 \tag{2.7}$$

This equation, together with Gauss' law

$$\nabla \cdot \varepsilon \mathbf{E} = \rho \tag{2.8}$$

form the foundations of the electrostatic regime. These two equations can also be expressed in integral form as:

$$\oint \mathbf{E} \cdot \mathbf{dl} = 0 \tag{2.9}$$

and

$$\int \varepsilon \mathbf{E} \cdot dA = \int \rho dV \tag{2.10}$$

© 2006 by Taylor & Francis Group, LLC

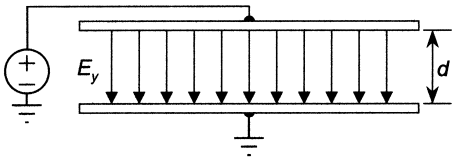


Figure 2.1 A simple system consisting of two parallel electrodes of area A separated by a distance d.

The curl-free electric field Eq. (2.7) can be expressed as the gradient of a scalar potential Φ:

$$\mathbf{E} = -\nabla\Phi \tag{2.11}$$

which can be integrated with respect to path length to yield the definition of the voltage difference between two points a and b:

$$V_{ab} = -\int_b^a \mathbf{E} \cdot \mathbf{dl} \tag{2.12}$$

Equation (2.12) applies in any geometry, but it becomes particularly simple for parallel-electrode geometry. For example, the two-electrode system of Fig. 2.1, with separation distance d, will produce a uniform electric field of magnitude

$$E_y = \frac{V}{d} \tag{2.13}$$

when energized to a voltage V. Applying Gauss' law to the inner surface of the either electrode yields a relationship between the surface charge ρ_s and E_y,

$$\varepsilon E_y = \rho_s \tag{2.14}$$

Here ρ_s has the units of coulombs per square meter, and ε is the dielectric permittivity of the medium between the electrodes. In other, more complex geometries, the solutions to Eqs. (2.9) and (2.10) take on different forms, as discussed in the next section.

2.3. DISCRETE AND DISTRIBUTED CAPACITANCE

When two conductors are connected to a voltage source, one will acquire positive charge and the other an equal magnitude of negative charge. The charge per unit voltage is called the *capacitance* of the electrode system and can be described by the relationship

$$C = \frac{Q}{V} \tag{2.15}$$

© 2006 by Taylor & Francis Group, LLC

Here $\pm Q$ are the magnitudes of the positive and negative charges, and V is the voltage applied to the conductors. It is easily shown that the capacitance between two parallel plane electrodes of area A and separation d is given approximately by

$$C = \frac{\varepsilon A}{d} \tag{2.16}$$

where ε is the permittivity of the material between the electrodes, and the approximation results because field enhancements, or "fringing effects," at the edges of the electrodes have been ignored. Although Eq. (2.16) is limited to planar electrodes, it illustrates the following basic form of the formula for capacitance in any geometry:

$$\text{Capacitance} = \frac{\text{permittivity} \times \text{area parameter}}{\text{length parameter}} \tag{2.17}$$

Table 2.1 provides a summary of the field, potential, and capacitance equations for energized electrodes in several different geometries.

2.4. DIELECTRIC PERMITTIVITY

The dielectric permittivity of a material describes its tendency to become internally polarized when subjected to an electric field. Permittivity in farads per meter can also can be expressed in fundamental units of coulombs per volt-meter (C/V·m). The *dielectric constant*, or *relative permittivity*, of a substance is defined as its permittivity normalized to ε_0, where $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m is the permittivity of free space. For reference purposes, relative permittivity values for several common materials are provided in Table 2.2. Note that no material has a permittivity smaller than ε_0.

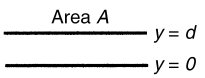
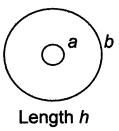
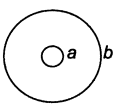
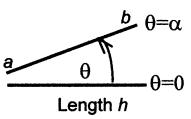
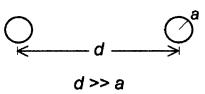
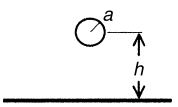
2.5. THE ORIGINS OF ELECTROSTATIC CHARGE

The source of electrostatic charge lies at the atomic level, where a nucleus having a fixed number of positive protons is surrounded by a cloud of orbiting electrons. The number of protons in the nucleus gives the atom its unique identity as an element. An individual atom is fundamentally charge neutral, but not all electrons are tightly bound to the nucleus. Some electrons, particularly those in outer orbitals, are easily removed from individual atoms. In conductors such as copper, aluminum, or gold, the outer electrons are weakly bound to the atom and are free to roam about the crystalline matrix that makes up the material. These free electrons can readily contribute to the flow of electricity. In insulators such as plastics, wood, glass, and ceramics, the outer electrons remain bound to individual atoms, and virtually none are free to contribute to the flow of electricity.

Electrostatic phenomena become important when an imbalance exists between positive and negative charges in some region of interest. Sometimes such an imbalance occurs due to the phenomenon of *contact electrification* [1–8]. When dissimilar materials come into contact and are then separated, one material tends to retain more electrons and become negatively charged, while the other gives up electrons and become positively charged. This contact electrification phenomenon, called *triboelectrification*, occurs at the

© 2006 by Taylor & Francis Group, LLC

Table 2.1 Field, Potential, and Capacitance Expressions for Various Electrode Geometries

Geometry	**E** field	Potential	Capacitance
Planar 	$E_y = \frac{V}{d}$	$\Phi = V\frac{y}{d}$	$C = \frac{\varepsilon A}{d}$
Cylindrical 	$E_r = \frac{V}{r\ln(b/a)}$	$\Phi = \frac{V}{\ln(b/a)}\ln\left(\frac{b}{r}\right)$	$C = \frac{2\pi\varepsilon h}{\ln(b/a)}$
Spherical 	$E_r = \frac{V}{r^2[1/a - 1/b]}$	$\Phi = V\frac{a}{r}\frac{(b-r)}{(b-a)}$	$C = 4\pi\varepsilon\frac{ba}{b-a}$
Wedge 	$E_\theta = \frac{V}{\alpha r}$	$\Phi = V\frac{\theta}{\alpha}$	$C = \frac{\varepsilon h}{\alpha}\ln\left(\frac{b}{a}\right)$
Parallel lines (at $\pm V$) 		$\Phi \approx \frac{2\pi\varepsilon V}{\ln(d/a)}\ln\left(\frac{r_1}{r_2}\right)$ r_1; r_2 = distances to lines	$C \approx \frac{\pi\varepsilon h}{\ln(d/a)}$
Wire to plane 			$C \approx \frac{2\pi\varepsilon}{\cosh^{-1}[(h+a)/a]}$ $h \gg a$

points of intimate material contact. The amount of charge transferred to any given contact point is related to the work function of the materials. The process is enhanced by friction which increases the net contact surface area. Charge separation occurs on both conductors and insulators, but in the former case it becomes significant only when at least one of the conductors is electrically isolated and able to retain the separated charge. This situation is commonly encountered, for example, in the handling of conducting powders. If neither conductor is isolated, an electrical pathway will exist between them, and the separated charges will flow together and neutralize one another. In the case of insulators, however, the separated charges cannot easily flow, and the surfaces of the separated objects remain charged. The widespread use of insulators such as plastics and ceramics in industry and manufacturing ensures that triboelectrification will occur in numerous situations. The pneumatic transport of insulating particles such as plastic pellets, petrochemicals, fertilizers, and grains are particularly susceptible to tribocharging.

© 2006 by Taylor & Francis Group, LLC

Table 2.2 Relative Permittivities of Various Materials

Air	1	Polycarbonate	~3.0
Alumina	8.8	Polyethylene	2.3
Barium titanate ($BaTiO_3$)	1200	Polyamide	~3.4–4.5
Borosilicate glass	4	Polystyrene	2.6
Carbon tetrachloride	2.2	Polyvinyl chloride	6.1
Epoxy	~3.4–3.7	Porcelain	~5–8
Ethanol	24	Quartz	3.8
Fused quartz (SiO_2)	3.9	Rubber	~2–4
Gallium arsenide	13.1	Selenium	6
Glass	~4–9	Silicon	11.9
Kevlar	~3.5–4.5	Silicon nitride	7.2
Methanol	33	Silicone	~3.2–4.7
Mylar	3.2	Sodium chloride	5.9
Neoprene	~4–6.7	Styrofoam	1.03
Nylon	~3.5–4.5	Teflon	2.1
Paper	~1.5–3	Water	~80
Paraffin	2.1	Wood (dry)	1.4–2.9
Plexiglas	2.8		

Table 2.3 The Triboelectric Series

POSITIVE	
Quartz	Copper
Silicone	Zinc
Glass	Gold
Wool	Polyester
Polymethyl methacrylate (Plexiglas)	Polystyrene
Salt (NaCl)	Natural rubber
Fur	Polyurethane
Silk	Polystyrene
Aluminum	Polyethylene
Cellulose acetate	Polypropylene
Cotton	Polyvinyl chloride
Steel	Silicon
Wood	Teflon
Hard rubber	NEGATIVE

Source: Compiled from several sources [9–13].

The relative propensity of materials to become charged following contact and separation has traditionally been summarized by the *triboelectric series* of Table 2.3. (*Tribo* is a Greek prefix meaning *frictional*.) After a contact-and-separation event, the material that is listed higher in the series will tend to become positively charged, while the one that is lower in the series will tend to become negatively charged. The vagueness of the phrase "will tend to" in the previous sentence is intentional. Despite the seemingly reliable order implied by the triboelectric series, the polarities of tribocharged materials often cannot be predicted reliably, particularly if the materials lie near each other in the series. This imprecision is evident in the various sources [9–13] cited in Table 2.3 that differ on the exact order of the series. Contact charging is an imprecise science that is driven by effects

© 2006 by Taylor & Francis Group, LLC

occurring on an atomic scale. The slightest trace of surface impurities or altered surface states can cause a material to deviate from the predictions implied by the triboelectric series. Two contact events that seem similar on the macroscopic level can yield entirely different results if they are dissimilar on the microscopic level. Thus contact and separation of *like* materials can sometimes lead to charging if the contacting surfaces are microscopically dissimilar. The triboelectric series of Table 2.3 should be viewed as a probabilistic prediction of polarity during multiple charge separation events. Only when two materials are located at extremes of the series can their polarities be predicted reliably following a contact-charging event.

2.6. WHEN IS "STATIC" CHARGE TRULY STATIC?

The term *static electricity* invokes an image of charge that cannot flow because it is held stationary by one or more insulators. The ability of charge to be static in fact does depend on the presence of an insulator to hold it in place. What materials can really be considered insulators, however, depends on one's point of view. Those who work with electrostatics know that the arrival of a cold, dry winter is synonymous with the onset of "static season," because electrostatic-related problems are exacerbated by a lack of humidity. When cold air enters a building and is warmed, its relative humidity declines noticeably. The tendency of hydroscopic surfaces to absorb moisture, thereby increasing their surface conductivities, is sharply curtailed, and the decay of triboelectric charges to ground over surface-conducting pathways is slowed dramatically. Regardless of humidity level, however, these conducting pathways always exist to some degree, even under the driest of conditions. Additionally, surface contaminants such as dust, oils, or residues can add to surface conduction, so that eventually all electrostatic charge finds its way back to ground. Thus, in most situations of practical relevance, no true insulator exists. In electrostatics, the definition of an insulator really depends on how long one is willing to wait. Stated succinctly, if one waits long enough, everything will look like a perfect conductor sooner or later. An important parameter associated with "static electricity" is its relaxation time constant—the time it takes for separated charges to recombine by flowing over conducting pathways. This relaxation time, be it measured in seconds, hours, or days, must always be compared to time intervals of interest in any given situation.

2.7. INDUCTION CHARGING

As discussed in the previous section, contact electrification can result in the separation of charge between two dissimilar materials. Another form of charge separation occurs when a voltage is applied between two conductors, for example the electrodes of a capacitor. Capacitive structures obey the relationship

$$Q = \pm CV \tag{2.18}$$

where the positive and negative charges appear on the surfaces of the opposing electrodes. The electrode which is at the higher potential will carry $+Q$; the electrode at the lower potential will carry $-Q$. The mode of charge separation inherent to capacitive structures is known as *inductive* charging. As Eq. (2.18) suggests, the magnitude of the inductively separated charge can be controlled by altering either C or V. This feature of induction

© 2006 by Taylor & Francis Group, LLC

charging lies in contrast to triboelectrification, where the degree of charge separation often depends more on chance than on mechanisms that can be controlled.

If a conductor charged by induction is subsequently disconnected from its source of voltage, the now electrically floating conductor will retain its acquired charge regardless of its position relative to other conductors. This mode of induction charging is used often in industry to charge atomized droplets of conducting liquids. The sequence of diagrams shown in Fig. 2.2 illustrates the process. The dispensed liquid becomes part the capacitive electrode as it emerges from the hollow tube and is charged by induction. As the droplet breaks off, it retains its charge, thereafter becoming a free, charged droplet. A droplet of a given size can be charged only to the maximum *Raleigh limit* [9,14,15]:

$$Q_{\max} = 8\pi\sqrt{\varepsilon_0 \gamma} R_p^{3/2} \tag{2.19}$$

Here γ is the liquid's surface tension and R_p the droplet radius. The Raleigh limit signifies the value at which self repulsion of the charge overcomes the surface tension holding the droplet together, causing the droplet to break up.

2.8. DIELECTRIC BREAKDOWN

Nature is fundamentally charge neutral, but when charges are separated by any mechanism, the maximum quantity of charge is limited by the phenomenon of *dielectric breakdown*. Dielectric breakdown occurs in solids, liquids and gases and is characterized by the maximum field magnitude that can be sustained before a field-stressed material loses its insulating properties.* When a solid is stressed by an electric field, imperfections

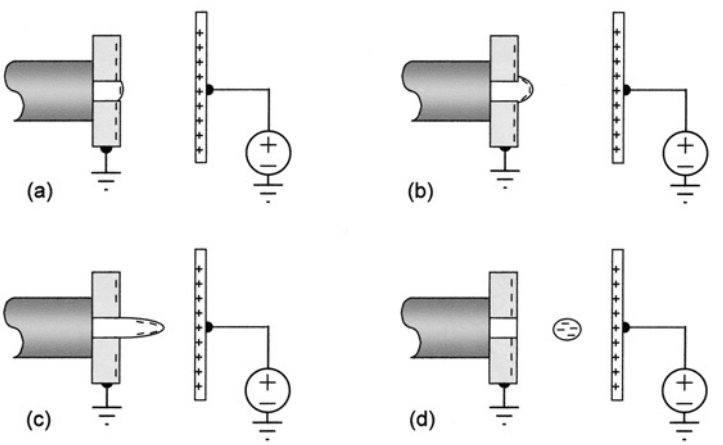


Figure 2.2 Charging a conducting liquid droplet by induction. As the droplet breaks off (d), it retains the charge induced on it by the opposing electrode.

*Breakdown in vacuum invariably occurs over the surfaces of insulating structures used to support opposing electrodes.

© 2006 by Taylor & Francis Group, LLC

or stray impurities can initiate a local discharge, which degrades the composition of the material. The process eventually extends completely through the material, leading to irreversible breakdown and the formation of a conducting bridge through which current can flow, often with dramatic results. In air and other gases, ever-present stray electrons (produced randomly, for example, by ionizing cosmic rays) will accelerate in an electric field, sometimes gaining sufficient energy between collisions to ionize neutral molecules, thereby liberating more electrons. If the field is of sufficient magnitude, the sequence of ensuing collisions can grow exponentially in a self-sustaining *avalanche* process. Once enough electrons have been liberated from their molecules, the gas becomes locally conducting, resulting in a spark discharge. This phenomenon is familiar to anyone who has walked across a carpet on a dry day and then touched a doorknob or light switch. The human body, having become electrified with excess charge, induces a strong electric field on the metal object as it is approached, ultimately resulting in the transfer of charge via a rapid, energetic spark. The most dramatic manifestation of this type of discharge is the phenomenon of atmospheric lightning.

A good rule of thumb is that air at standard temperature and pressure will break down at a field magnitude of about 30 kV/cm (i.e., 3 MV/m or 3×10^6 V/m). This number increases substantially for small air gaps of 50 μm or less because the gap distance approaches the mean free path for collisions, and fewer ionizing events take place. Hence a larger field is required to cause enough ionization to initiate an avalanche breakdown. This phenomena, known as the *Paschen effect*, results in a breakdown-field versus gap-distance curve such as the one shown in Fig. 2.3 [9,10,12,18,19]. The Paschen effect is critical to the operation of micro-electromechanical systems, or MEMS, because fields in excess of 30 kV/cm are required to produce the forces needed to move structural elements made from silicon or other materials.

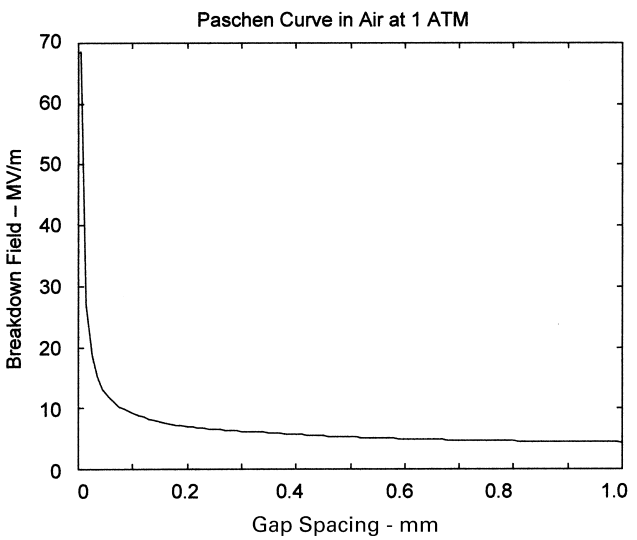


Figure 2.3 Paschen breakdown field vs. gap spacing for air at 1 atmosphere. For large gap spacings, the curve is asymptotic to 3×10^6 V/m.

© 2006 by Taylor & Francis Group, LLC

2.9. CORONA DISCHARGE

One of the more common methods for intentionally producing electrostatic charge involves the phenomenon of *corona discharge*. Corona is a partial breakdown that occurs when two electrodes, one sharp and the other much less so, are energized by a voltage source. In such a configuration, the electric field around the sharp electrode is greatly enhanced. At some critical level of voltage, called the *onset voltage*, the field near the sharp electrode exceeds the dielectric breakdown strength of the gas, typically air. This localized breakdown produces free electrons and positive ions via the avalanche process. In the remainder of the electrode space, however, the field is substantially weaker, and no ionization takes place. Thus the breakdown that occurs near the stressed electrode provides a source of ions, but no spark discharge occurs. If the stressed electrode is positive, the positive ions will be repelled from it, providing an abundant source of positive ions. If the stressed electrode is negative, the free electrons will be repelled from it but will quickly attach to neutral molecules upon leaving the high field region, thereby forming negative ions. The phenomenon of corona is illustrated graphically in Fig. 2.4 for a positive source electrode.

For either ion polarity, and in most electrode configurations, the relationship between applied voltage and the resulting corona current follows an equation of the form $i_C = gV(V - V_C)$, where V_C is the critical onset voltage of the electrode system and g is a constant. The values of g and V_C will depend on many factors, including electrode geometry, spacing, radii of curvature, and surface roughness, as well as on ion mobility, air temperature, and air pressure. One must generally determine g and V_C empirically, but in coaxial geometry this relationship can be solved analytically [18]. The result is a complex formula, but for small currents, the equation for cylindrical geometry can be approximated by

$$i_L = \frac{4\pi\varepsilon_0 \kappa V(V - V_C)}{b^2 \ln(b/a)} \tag{2.20}$$

Here i_L is the current per unit axial length, b and a are the outer and inner coaxial radii, respectively, and κ is the ion mobility (about $2.2 \times 10^{-4}\,\mathrm{m^2/V{\cdot}s}$ for air at standard temperature and pressure). As the applied voltage V is increased, corona will first occur at the corona onset voltage V_C. For coaxial electrodes with an air dielectric, V_C is equal to

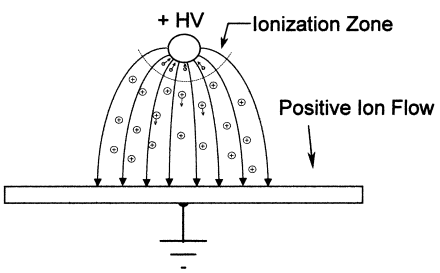


Figure 2.4 Basic mechanism of corona discharge near a highly stressed electrode. Positive corona is shown; a similar situation exists for negative corona.

© 2006 by Taylor & Francis Group, LLC

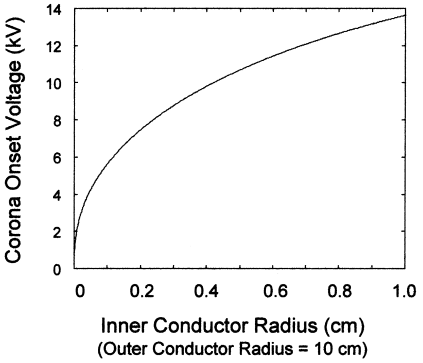


Figure 2.5 Plot of corona onset voltage V_c vs. inner conductor radius a for coaxial electrodes with 10-cm outer conductor radius.

the voltage at which the electric field on the surface of the inner electrode first reaches the value given by Peek's equation [18,20]:

$$E_{\text{peek}} = mE_{\text{bk}}\left(1 + \frac{0.0308}{\sqrt{a}}\right) \tag{2.21}$$

Here $E_{\text{bk}} = 3 \times 10^6$ V/m is the breakdown strength of air under uniform field conditions, a is the inner conductor radius in meters, m is an empirical surface roughness factor, and standard temperature and pressure are assumed. Note that E_{peek} will always be larger than the breakdown field E_{bk}. Peek's equation describes the field that must be established at the inner conductor surface before local breakdown (corona) can occur. The equation is also approximately valid for parallel-wire lines. For smooth conductors $m = 1$, and for rough surfaces $m = 0.8$.

In a coaxial system, the electric field magnitude at the inner radius a is given by

$$E(r) = \frac{V}{a\ln(b/a)} \tag{2.22}$$

hence the corona onset voltage becomes

$$V_C = E_{\text{peek}}\, a \ln\left(\frac{b}{a}\right) = E_{\text{bk}}\left(1 + \frac{0.0308}{\sqrt{a}}\right) a \ln\left(\frac{b}{a}\right) \tag{2.23}$$

A plot of V_C versus a for the case $b = 10$ cm is shown in Fig. 2.5.

2.10. CHARGES AND FORCE

The electrostatic force f_{12} between two charges q_1 and q_2 separated by a distance r is governed by Coulomb's law, a fundamental principle of physics:

$$f_{12} = \frac{q_1 q_2}{4\pi\varepsilon r^2} \tag{2.24}$$

© 2006 by Taylor & Francis Group, LLC

The direction of this force is parallel to a line between the charges. All other force relationships in electrostatics derive from Coulomb's law. If a collection of charges produces a net electric field **E**, it is easily shown by integration that the collective force exerted on a solitary charge q by all the other charges becomes just $q\mathbf{E}$. This simple relationship comprises the electric field term in the Lorentz force law of electromagnetics:

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{2.25}$$

In many practical situations in electrostatics, one is interested in the forces on conductors and insulators upon which charges reside. Numerous mathematical methods exist for predicting such forces, including the *force-energy* method, the *boundary element* method, and the *Maxwell stress tensor* [21–24]. Of these three methods, the force-energy method is the one most easily understood from basic principles and the most practical to use in many situations. The analysis that follows represents an abridged derivation using the force-energy method.

We first consider a constant-charge system in which two objects carrying fixed charges experience a net force F_Q (as yet unknown). One such hypothetical system is illustrated in Fig. 2.6. If one of the objects is displaced against F_Q by an incremental distance dx relative to the other object, then the mechanical work dW_m performed on the displaced object will be $F_Q\,dx$. Because the objects and their fixed charges are electrically isolated, the work transferred to the displaced body must increase the energy stored in the system. The stored electrostatic energy W_e thus will be augmented by dW_m, from which it follows that

$$F_Q = \frac{dW_m}{dx} \tag{2.26}$$

As an example of this principle, consider the parallel-electrode structure of Fig. 2.7, for which the capacitance is given by

$$C = \frac{\varepsilon A}{x} \tag{2.27}$$

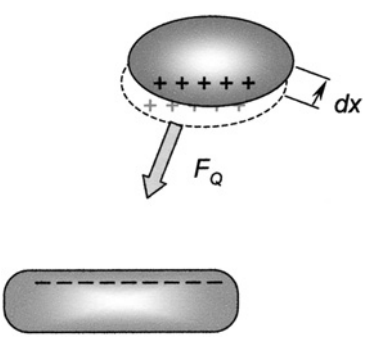


Figure 2.6 One charged object is displaced relative to another. The increment of work added to the system is equal the electrostatic force F_Q times the displacement dx.

© 2006 by Taylor & Francis Group, LLC

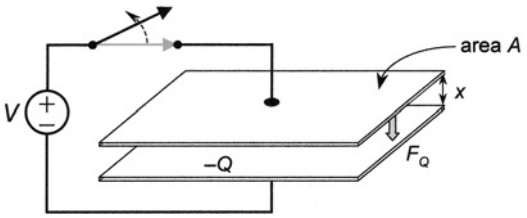


Figure 2.7 Parallel electrodes are energized by a voltage source that is subsequently disconnected. Fixed charges $\pm Q$ remain on the electrodes.

If the electrodes are precharged, then disconnected from their source of voltage, the charge will thereafter remain constant. The stored electrical energy can then be expressed as [24]

$$W_e = \frac{Q^2}{2C} \tag{2.28}$$

The force between the electrodes can be found by taking the x derivative of this equation:

$$F_Q = \frac{dW_e}{dx} = \frac{Q^2}{2}\frac{d}{dx}\left(\frac{x}{\varepsilon A}\right) = \frac{Q^2}{2\varepsilon A} \tag{2.29}$$

Equation (2.29) also describes the force between two insulating surfaces of area A that carry uniform surface charge densities $\rho_s = \pm Q/A$.

It is readily shown [21–24] that applying the energy method to two conductors left connected to the energizing voltage V yields a similar force equation:

$$F_V = \frac{dW_e}{dx} \tag{2.30}$$

Here W_e is the stored electric energy expressed as $1/2CV^2$. When this formula is applied to a system in which voltage, not charge, is constrained, the force it predicts will always be attractive.

Equation (2.30) is readily applied to the parallel-electrode structure of Fig. 2.7 with the switch closed. The force between the conductors becomes

$$\frac{dW_e}{dx} = \frac{V^2}{2}\frac{d}{dx}\left(\frac{\varepsilon A}{x}\right) = -\frac{\varepsilon A V^2}{2x^2} \tag{2.31}$$

This force is inversely proportional to the square of the separation distance x.

2.11. PARTICLE CHARGING IN AIR

Many electrostatic processes use the coulomb force to influence the transport of charged airborne particles. Examples include electrostatic paint spraying [10,16], electrostatic

© 2006 by Taylor & Francis Group, LLC

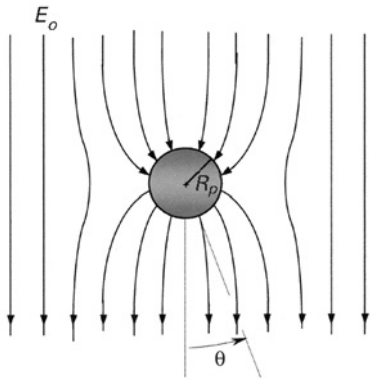


Figure 2.8 Conducting sphere distorts an otherwise uniform electric field. The field components are given by Eq. (2.32). If the source of the field produces ions, the latter will follow the field lines to the particle surface.

powder coating, electrostatic crop spraying [25,26], electrostatic drug delivery, and electrostatic precipitation. These processes are described later in this chapter. Airborne particles are sometimes charged by induction, requiring that initial contact be made with a conducting electrode. In other processes, particles are charged by ions in the presence of an electric field.

In this section, we examine the latter process in more detail. To a first approximation, many airborne particles can be treated as conducting spheres—an assumption that greatly simplifies the equations governing particle charging. The approximation requires that the particle have a shape free from prominent asymmetries and also that the intrinsic charging time of the particle, given by the ratio ε/σ of the particle's permittivity to conductivity, be much shorter than other time scales of interest. Suppose that an uncharged particle of radius R_p is situated in a uniform, downward-pointing electric field $\boldsymbol{E}_o$, as depicted in Fig. 2.8. A "uniform field" in this case is one that does not change spatially over the scale of at least several particle radii. Further suppose that a uniform, homogeneous source of unipolar ions is produced by the system and carried toward the particle by the electric field. These ions might be produced, for example, by some form of corona discharge. If we assume the ion density to be small enough such that space-charge perturbation of the field is negligible, the electric field components in the neighborhood of the particle become:

$$E_r = E_o\left(1 + \frac{2R_p^3}{r^3}\right)\cos\theta + \frac{Q}{4\pi\varepsilon_o r^2}$$

and

$$E_\theta = E_o\left(\frac{R_p^3}{r^3} - 1\right)\sin\theta \tag{2.32}$$

with $E_\varphi = 0$. Here Q represents any charge that the conducting particle may carry. If Q is positive, the second term in the equation for E_r adds a uniform radial component that points outward.

© 2006 by Taylor & Francis Group, LLC

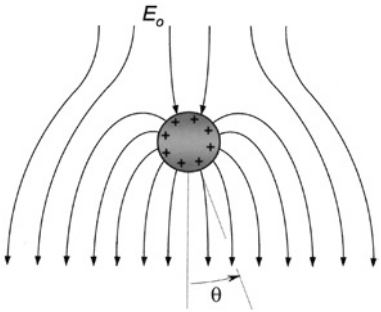


Figure 2.9 As the particle collects charge Q, the field lines are increasingly excluded from the particle surface.

Figure 2.8 shows the field pattern for the case $Q = 0$. Note that **E** is everywhere perpendicular to the particle surface, where $E_\theta = 0$. Ions will be transported to the surface of the particle by the field, thereby increasing the magnitude of Q. If the ions are positive, only field lines leading *into* the particle will contribute to its charging. Field lines that originate from the surface of the particle cannot carry ions, because no source of ions exists there. As charge accumulates on the particle and the second term for E_r in Eq. (2.32) becomes larger, the field pattern for $Q \neq 0$ takes the form shown in Fig. 2.9. The reduction in magnitude of the inward-pointing field lines restricts the flow of ions to the particle surface. When $Q/4\pi\varepsilon_0 r^2$ in Eq. (2.32) becomes equal to the factor $E_o(1 + 2R_p^3/r^3)$ at $r = R_p$, all field lines will originate from the particle itself, so that further ion charging of the particle will cease. Under this condition, E_r at $\theta = 180^\circ$ and $r = R_p$ becomes zero. The charge limit Q_{sat} can thus be found by setting E_r in Eq. (2.32) to zero, yielding

$$\frac{Q_{\text{sat}}}{4\pi\varepsilon_0 R_p^2} = 3E_o \tag{2.33}$$

or

$$Q_{\text{sat}} = 12\pi\varepsilon_0 R_p^2 E_o \tag{2.34}$$

The value given by Eq. (2.34) is called the *saturation charge* of the particle, or sometimes the Pauthenier limit [27]. It represents the maximum charge that the particle can hold. For a 100-μm particle situated in a 100-kV/m field, for example, the saturation charge calculated from Eq. (2.34) becomes 0.33 pC.

Note that Q_{sat} increases with particle radius and the ambient field E_o, but it is not dependent on ion mobility or ion density. These latter quantities affect only the *rate* of particle charging [15,24].

For $Q < Q_{\text{sat}}$, it can be shown via surface integration of the field equation, Eq. (2.32), that the ion current to the particle is given by

$$i_Q = \frac{dQ}{dt} = 3\pi R_p^2 E_o N q_{\text{ion}} \kappa \left(1 - \frac{Q}{Q_{\text{sat}}}\right)^2 \tag{2.35}$$

© 2006 by Taylor & Francis Group, LLC

where N is the ambient ion density, q_{ion} the ion charge, and κ the ion mobility. Solving this differential equation results in an expression for Q as a function of time:

$$Q(t) = Q_{\text{sat}} \frac{t/\tau}{1 + t/\tau} \tag{2.36}$$

This hyperbolic charging equation is governed by the time constant $\tau = 4\varepsilon_0/Nq_{\text{ion}}\,\kappa$. For the typical values $N = 10^{15}$ ions/m^3 and $\kappa = 2 \times 10^{-4}$ m^2/V·s for singly charged ions in air, particle charging will be governed by the hyperbolic charging time constant $\tau = 1.1$ ms. Note that this latter value is independent of particle radius and electric field magnitude.

2.12. CHARGED PARTICLE MOTION

A charged, airborne particle will experience two principal forces: electrostatic and aerodynamic. The former will be given by

$$\mathbf{F}_{\text{elec}} = Q\mathbf{E} \tag{2.37}$$

where Q is the particle charge, while the latter will be given by the Stokes' drag equation [9,15]:

$$\mathbf{F}_{\text{drag}} = -6\pi\eta R_p(\mathbf{U}_{\text{p}} - \mathbf{U}_{\text{air}}) \tag{2.38}$$

Here $\mathbf{U}_{\text{p}}$ is the particle velocity, $\mathbf{U}_{\text{air}}$ the ambient air velocity (if any), and η the kinematic viscosity of air. At standard temperature and pressure, $\eta = 1.8 \times 10^{-5}$ N·s/m^2 [9]. Equation (2.38) is valid for particles in the approximate size range 0.5 to 25 μm, for which inertia can usually be ignored. For smaller particles, Brownian motion becomes the dominant mechanical force, whereas for particles larger than about 25 μm, the Reynolds number for typical values of $\mathbf{U}_{\text{p}}$ approaches unity and the Stokes' drag limit no longer applies.

The balance between $\mathbf{F}_{\text{elec}}$ and $\mathbf{F}_{\text{drag}}$ determines the net particle velocity:

$$\mathbf{U}_{\text{p}} = \mathbf{U}_{\text{air}} + \frac{Q}{6\pi\eta R_p}\text{E} \tag{2.39}$$

The quantity $Q/6\pi\eta R_p$, called the particle *mobility*, describes the added particle velocity per unit electric field. The mobility has the units of m^2/V · s.

2.13. ELECTROSTATIC COATING

Electrostatic methods are widely used in industry to produce coatings of excellent quality. Electrostatic-assisted spraying techniques can be used for water or petroleum-based paints as well as curable powder coatings, surface lacquers, and numerous chemical substrates. In electrostatic paint spraying, microscopic droplets charged by induction are driven directly to the surface of the work piece by an applied electric field. In power coating methods, dry particles of heat-cured epoxies or other polymers are first charged, then forced to the surface of the work piece by electrostatic forces. Similar spraying techniques are used to

© 2006 by Taylor & Francis Group, LLC

coat crops with minimal pesticides [17, 25, 26]. In one system, electrostatic methods are even used to spray tanning solution or decontamination chemicals on the human body. Electrostatic methods substantially reduce the volume of wasted coating material, because particles or droplets are forced directly to the coated surface, and only small amounts miss the target to become wasted product.

2.14. ELECTROSTATIC PAINT SPRAYING

The basic form of an electrostatic paint spray system is illustrated in Fig 2.10. Paint is atomized from a pressurized nozzle that is also held at a high electric potential relative to ground. Voltages in the range 50 kV to 100 kV are typical for this application. As paint is extruded at the nozzle outlet, it becomes part of the electrode system, and charge is induced on the surface of the liquid jet. This charge will have the same polarity as the energized nozzle. As each droplet breaks off and becomes atomized, it carries with it its induced charge and can thereafter be driven by the electric field to the work piece. Very uniform charge-to-mass ratios Q/m can be produced in this way, leading to a more uniform coating compared to nonelectrostatic atomization methods. The technique works best if the targeted object is a good conductor (e.g., the fender of an automobile), because the electric field emanating from the nozzle must terminate primarily on the work piece surface if an efficient coating process is to be realized.

Once a droplet breaks away from the nozzle, its trajectory will be determined by a balance between electrostatic forces and viscous drag, as summarized by Eq. (2.39). If the ambient air velocity $\mathbf{U}_{\text{air}}$ is zero, this equation becomes

$$\mathbf{U}_{\text{p}} = \frac{Q}{6\pi\eta R_p}\mathbf{E} \tag{2.40}$$

Of most interest is the droplet velocity when it impacts the work surface. Determining its value requires knowledge of the electric field at the work surface, but in complex

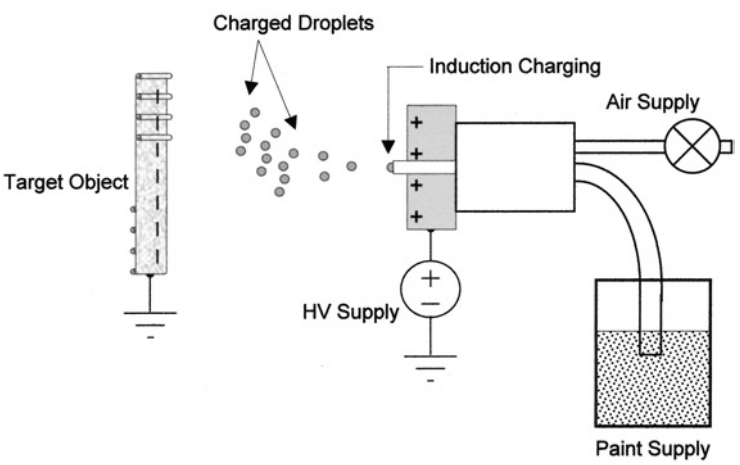


Figure 2.10 Basic electrostatic spray system. Droplets are charged by induction as they exit the atomization nozzle.

© 2006 by Taylor & Francis Group, LLC

geometries, analytical solutions are seldom possible. Estimation or empirical measurement using a field mill (see Sec. 2.18) is usually required. Velocities in the range 0.1 to 100 m/s are common in electrostatic painting operations [9–16].

Note that the particle radius in Eq. (2.40) can be expressed in terms of the droplet mass, given by

$$M = \frac{4\pi R_p^3 \gamma}{3} \tag{2.41}$$

where is the mass density of the liquid. The velocity equation, Eq. (2.40), can thus be written as

$$\mathbf{U}_{\mathrm{p}} = \frac{2R_p^2 \gamma}{9\eta} \frac{Q}{M} \mathbf{E} \tag{2.42}$$

This form of the equation illustrates the significance of the *charge-to-mass ratio* of the droplet. For a given electric field magnitude, the droplet velocity will be proportional to Q/M. Because Q has a maximum value determined by either the Raleigh limit of Eq. (2.19) or the saturation charge limit of Eq. (2.34), Eq. (2.42) will be limited as well. For a 100-μm droplet of unity density charged to its saturation limit in a 100 kV/m field, the impact velocity becomes about 0.1 m/s.

2.15. ELECTROPHOTOGRAPHY

The "simple" copy machine has become common in everyday life, but in reality, this machine is far from simple. The copier provides a good example of how electrostatics can be used to transfer particles between surfaces. The transfer process, first invented by Chester Carlson around 1939 [10], is also known as *electrophotography*, or sometimes *xerography*. Although the inner workings of a copy machine are complex [28], its basic features can be understood from the simplified diagram of Fig. 2.11. A thin photosensitive layer is deposited over a grounded surface, usually in the form of a rotating drum. The photosensitive material has the property that it remains an insulator in the dark but becomes partially conducting when exposed to light.

In the first step, the photoconductor is charged by ions from a corona source. This device, sometimes called a *corotron* [21], is scanned just over the surface of the photoconductor, allowing ions to migrate and stick to the photoconductor surface. These deposited charges are strongly attracted to their image charges in the underlying ground layer, but because the dark photoconductor is an insulator, the charges cannot move toward each other, but instead remain fixed in place.

Next, light projected from the image to be reproduced is focused on the photoconductor surface. The regions of the image corresponding to black remain insulating, while the white areas are exposed to light and become conducting. The charge deposited over these latter regions flows through the photoconductor to the ground plane, thereby discharging the photoconductor. The remaining electrostatic pattern on the drum is called a *latent image*.

The photoconductor is next exposed to toner particles that have been charged, usually by triboelectrification, to a polarity opposite that of the latent image. Some field

© 2006 by Taylor & Francis Group, LLC

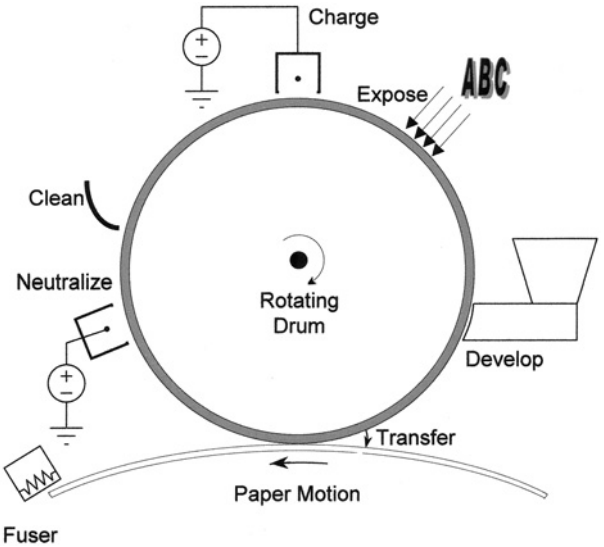


Figure 2.11 Basic elements of an electrostatic photocopier. As the light-sensitive drum rotates, it is charged, exposed to the image, dusted with toner, brought into contact with the paper, then discharged, and cleaned. The imprinted page passes through a fuser which melts the toner into the paper.

lines from the latent image extend above the surface and are of sufficient magnitude to capture and hold the charged toner particles. The latent image is thus transformed into a real image in the form of deposited toner particles.

In the next step of the process, image on the toner-coated drum is transferred to paper. The paper, backed by its own ground plane, is brought in proximity to the photoconductor surface. If the parameters are correctly chosen, the toner particles will be preferentially attracted to the paper and will jump from the photoconductor to the paper surface. The paper is then run through a high-temperature fuser which melts the toner particles into the paper.

2.16. ELECTROSTATIC PRECIPITATION

Electrostatic precipitation is used to remove airborne pollutants in the form of smoke, dust, fumes, atomized droplets, and other airborne particles from streams of moving gas [29–34]. Electrostatic precipitators provide a low cost method for removing particles of diameter 10 μm or smaller. They are often found in electric power plants, which must meet stringent air quality standards. Other applications include the cleaning of gas streams from boilers, smelting plants, blast furnaces, cement factories, and the air handling systems of large buildings. Electrostatic precipitators are also found on a smaller scale in room air cleaners, smoke abatement systems for restaurants and bars, and air cleaning systems in restaurants and hospitals (e.g., for reducing cigarette smoke or airborne bacteria). Electrostatic precipitators provide an alternative to bag house filters which operate like large vacuum-cleaner bags that filter pollutants from flowing gas.

© 2006 by Taylor & Francis Group, LLC

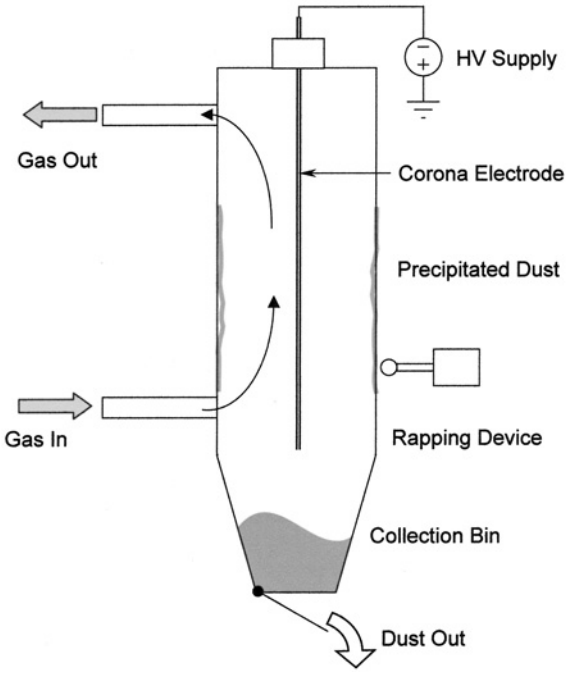


Figure 2.12 Schematic diagram of a single-stage, cylindrical electrostatic precipitator. Dust-laden air enters at the bottom of the stack; clean air exits the top. Negative corona current charges the particles, which then precipitate on the chamber walls. Mechanical "rapping" is used to help dislodge the dust to the collection bin.

The pressure drop across a bag-house filtration system can be very large, hence smaller pressure drop is one principal advantage of electrostatic precipitator systems. Another advantage of an electrostatic precipitator is its lower power consumption compared to a bag-house system, because less air handling equipment is required. The overall pressure drop in a large, industrial-scale electrostatic precipitator, for which the gas flow rate may exceed $1000\,m^3$/min, is typically less than 10 mm H_2O from source to exit [32].

The basic elements of a precipitator system are shown in Fig. 2.12. The particle-laden gas stream flows through a collection of corona electrodes mounted inside a rigid duct. The corona electrodes can be thin, parallel wires suspended on insulators, or a series of sharp points facing the duct walls. As discussed in Sec. 2.9, corona current will flow once the applied voltage exceeds the critical onset value expressed by Peek's formula, Eq. (2.21). In a large industrial precipitator, this onset voltage might be in the tens-of-kilovolts range, while the onset voltage in a small scale room precipitator is usually below 10 kV. It is difficult to achieve stable corona discharge below about 5 kV because the small gap sizes required to achieve Peek's field often lead to complete spark breakdown across the electrode gap.

The electrodes in an electrostatic precipitator serve two functions. The corona discharge produces a steady stream of ions which charge the airborne particles via the ion-impact charging mechanism described in Sec. 2.11. The charged particles then experience a transverse coulomb force $q\mathbf{E}$ and migrate toward the walls of the duct where they are collected and later removed by one of several cleaning methods. These methods include

© 2006 by Taylor & Francis Group, LLC

periodic washing of the duct walls, mechanical rapping to cause the particles to fall into a collection bin, and replacement of the duct's inner lining. This last method is usually reserved for small, bench-top systems.

Although most airborne particles will be neither spherical nor perfectly conducting, the model of Eqs. (2.32)–(2.36) often provides a reasonable estimate of particle charging dynamics. One important requirement is that the particles have enough residence time in the corona-ion flux to become charged to saturation and to precipitate on the collection walls of the duct.

Two problems of concern in the design of electrostatic precipitators include gradient force motion of dielectric or conducting particles, and a phenomenon known as *back ionization*. Gradient force, which is independent of particle charge, occurs whenever a particle is situated in an electric field whose magnitude changes with position, that is, when $\nabla|\mathbf{E}| \neq 0$. This phenomenon is illustrated schematically in Fig. 2.13 for a conducting, spherical particle. The free electrons inside the particle migrate toward the left and leave positive charge to the right, thereby forming a dipole moment. The electric field in Fig. 2.13a is stronger on the right side of the particle, hence the positive end of the dipole experiences a stronger force than does the negative end, leading to a net force to the right. In Fig. 2.13b, the field gradient and force direction, but not the orientation of the dipole, are reversed. For the simple system of Fig. 2.13, the force can be expressed by the one-dimensional spatial derivative of the field:

$$F_x = (qd)\frac{dE_x}{dx} \tag{2.43}$$

In three-dimensional vector notation, the dipole moment is usually expressed as

$$\mathbf{p} = q\mathbf{d}$$

where $\mathbf{d}$ is a vector pointing from the negative charge of the dipole to its positive charge. Hence in three dimensions, Eq. (2.43) becomes

$$\mathbf{F} = (\mathbf{p} \cdot \nabla)\mathbf{E} \tag{2.44}$$

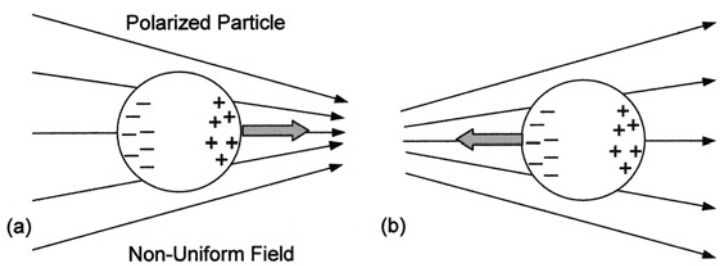


Figure 2.13 A conducting or dielectric particle in a nonuniform field. The particle is polarized, pulling the positive end in the direction of the field lines and the negative end against them. The net force on the particle will be toward the side that experiences the stronger field magnitude. (a) Positive force dominates; (b) negative force dominates.

© 2006 by Taylor & Francis Group, LLC

Equation (2.44) also applies in the more general case of a dielectric particle, where the force density **f** is expressed in terms of the polarization vector $\mathbf{P} = n\mathbf{p}$:

$$\mathbf{f} = (\mathbf{P} \cdot \nabla)\mathbf{E} \tag{2.45}$$

where n is the number of polarized dipoles per unit volume.

In the corona electrode configuration of an electrostatic precipitator, the field gradient is most pronounced near the corona-producing electrode. Here the gradient force can exceed the coulomb force in magnitude and cause pollutant particles to migrate toward, and deposit on, the high voltage electrode rather than on the collector plate. This phenomenon reduces the efficiency of the precipitator but can be avoided by ensuring that the particles acquire saturation charge quickly as they flow through the duct.

The second problem in precipitation, called *back ionization*, or sometimes *back corona* [32,33], occurs when the precipitated particles have high resistivity. The corona current passing through the built-up layer on its way to the duct walls can raise the surface potential of the layer. If this surface potential exceeds the breakdown strength of air, a discharge occurs in the layer, liberating electrons and producing positive ions. These ions migrate toward the negative electrode and tend to neutralize the pollutant particles. This process can greatly reduce the collection efficiency of the precipitator.

2.17. FIELD AND CHARGE MEASUREMENT

The ability to measure electrostatic fields and charge is important in many scientific and engineering disciplines. Measuring these quantities usually requires specialized instrumentation, because a standard voltmeter is useful in only a limited set of circumstances. For example, if one attempts to measure the potential of a charged, electrically isolated conductor with a voltmeter, as in Fig. 2.14, the internal impedance of the meter will fix the conductor potential at zero and allow its charge to flow to ground, thereby obscuring the original quantity to be measured. A standard voltmeter is altogether useless for measuring the potential of a charged *insulator*, because a voltmeter requires that some current, however small, be drawn from the point of measurement. Moreover, the surface of a charged insulator need not be an equipotential; hence the concept of voltage becomes somewhat muddied.

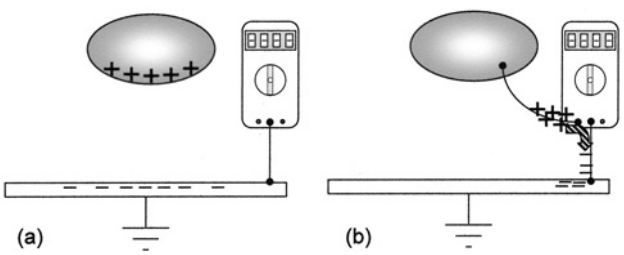


Figure 2.14 Attempting to measure the voltage of a charged, isolated conductor (a) results in a discharged object of zero potential (b).

© 2006 by Taylor & Francis Group, LLC

2.18. ELECTROSTATIC FIELD MILL

Numerous devices have been developed to measure electrostatic fields and voltages, including force sensors [35] and high-impedance solid-state sensors [35–38], but the most prevalent for measuring electrostatic fields has been the variable capacitance field mill [9,10,39,40]. The term *field mill* is used here in its broadest sense to describe any electrostatic field measuring device that relies on mechanical motion to vary the capacitance between the sensor and the source of the field. The variable aperture variety is prevalent in atmospheric science, electric power measurements, and some laboratory instruments, while the vibrating capacitor version can be found in numerous laboratory instruments.

The motivation for the variable aperture field mill comes from the boundary condition for an electric field incident upon a grounded, conducting electrode:

$$\varepsilon E = \rho_s \tag{2.46}$$

or

$$E = \frac{\rho_s}{\varepsilon} \tag{2.47}$$

where ρ_s is the surface chanrge density. A variable aperture field mill modulates the exposed area of a sensing electrode, so that the current flowing to the electrode becomes

$$i = \frac{dQ}{dt} = \frac{d\rho_s A}{dt} = \varepsilon E \frac{dA}{dt} \tag{2.48}$$

For a time-varying, periodic $A(t)$, the peak current magnitude will be proportional to the electric field incident upon the field mill.

One type of variable aperture field mill is depicted in Fig. 2.15. A vibrating vane periodically blocks the underlying sense electrode from the incident field, thereby causing

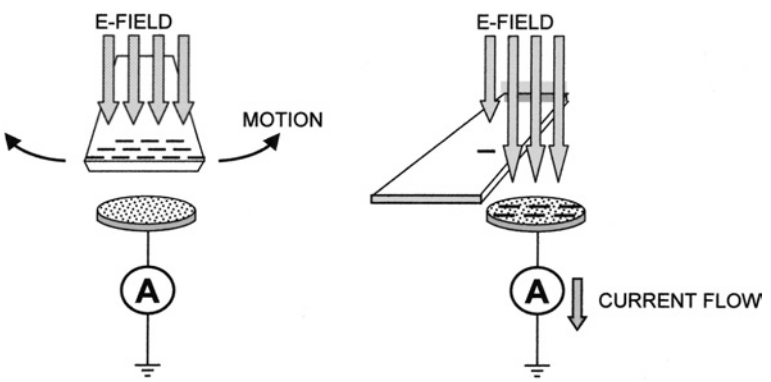


Figure 2.15 Simplified rendition of the variable-aperture field mill.

© 2006 by Taylor & Francis Group, LLC

the induced charge to change periodically. If the exposed area varies sinusoidally as

$$A = A_o \frac{1 + \sin \omega t}{2} \tag{2.49}$$

then the peak current to the sensing electrode will be given by

$$i_{\text{peak}} = \omega \varepsilon E \frac{A_o}{2} \tag{2.50}$$

If the electric field strength varies spatially on a scale comparable to the span of the aperture, then the field mill will respond to the spatial average of the incident field taken over the aperture area. Fields with small-scale spatial variations are found in several industrial, biological, and micromechanical applications. Aperture diameters as small as 0.5 mm are practical and may be found inside the probes of commercially available field meters and noncontacting voltmeters.

2.19. NONCONTACTING VOLTMETER

The field mill described in the previous section is important to an instrument known as the feedback-null surface potential monitor, or *noncontacting voltmeter* [9,10,41–43]. Commercial versions of these instruments are standard equipment in most electrostatics laboratories. The most salient feature of this measurement method is that surface potentials can be measured without physical contact. The basic operating principle of the meter is illustrated in Fig. 2.16. A small field mill is mounted on the end of a hand-held probe, but its outer housing is *not* connected to ground. The output signal of the field-mill feeds a phase-sensitive detection circuit and high-voltage amplifier. The output of the latter is connected back to the probe housing, thereby forming a negative feedback loop. When the probe encounters an object at nonzero potential, the detected field signal, amplified by the high-voltage amplifier, raises the potential of the probe until the field incident on the probe approaches zero. This concept is illustrated in Fig. 2.17. The feedback loop attains equilibrium when the probe body is raised to the same potential as the surface being measured, resulting in only a small residual field at the probe aperture. The residual signal in this "null-field" condition can be made arbitrarily small by increasing the gain of the high-voltage amplifier. Under equilibrium feedback conditions,

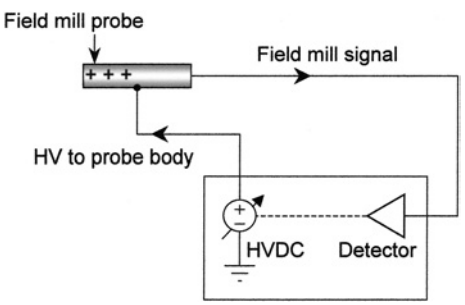


Figure 2.16 Basic structure of a noncontacting voltmeter.

© 2006 by Taylor & Francis Group, LLC

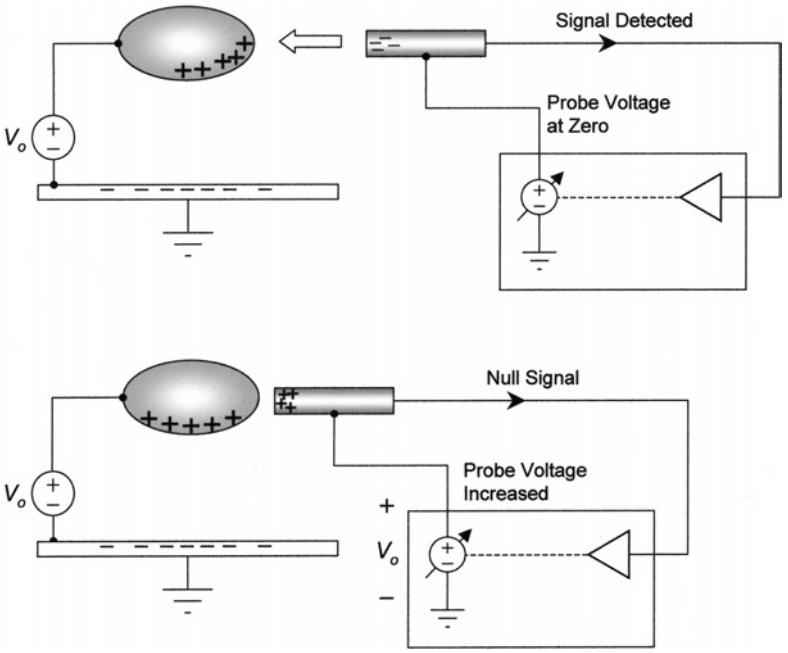


Figure 2.17 The noncontacting voltmeter in operation. Top: Probe approaches charged object to be measured. Bottom: Probe potential is raised until the field it measures is zero (null signal condition).

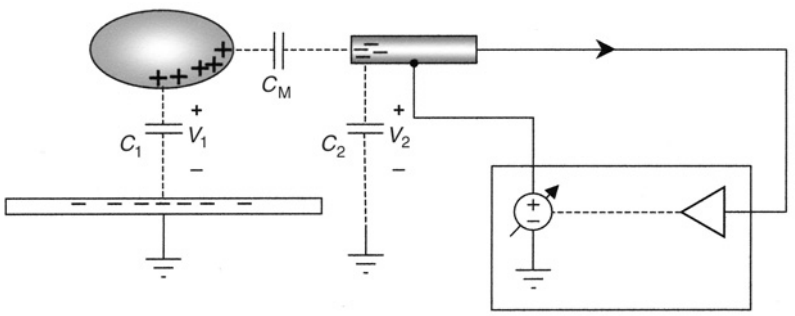


Figure 2.18 Using a noncontacting voltmeter to measure a charged, electrically isolated conductor.

the high voltage on the probe body, monitored using any suitable metering circuit, provides a measure of the surface potential. If the surface potential varies spatially, the meter output will reflect the spatial average encountered by the probe's aperture. The measuring range of the instrument is determined by the positive and negative saturation limits of the high-voltage amplifier. Values up to a few kilovolts (positive and negative) are typical for most commercial instruments.

When a noncontacting voltmeter reads the surface of a conductor connected to a fixed voltage source, the reading is unambiguous. If the probe approaches a floating conductor, the situation can be modeled by the two-body capacitance system of Fig. 2.18.

© 2006 by Taylor & Francis Group, LLC

In this diagram, C_1 and C_2 denote the capacitances to ground of the conductor and probe, respectively, and C_M represents their mutual capacitance. The charge Q_1 on the conductor will be given by [44]

$$Q_1 = C_1 V_1 + C_{\mathrm{M}}(V_1 - V_2) \tag{2.51}$$

The feedback loop of the meter will raise the potential of the probe until $V_2 = V_1$, so that Eq. (2.51) becomes

$$V_1 = \frac{Q_1}{C_1} \tag{2.52}$$

This unambiguous result reflects the potential of the floating conductor with the probe absent.

One of the more common uses of noncontacting voltmeters involves the measurement of charge on insulating surfaces. If surface charge on an insulating layer is tightly coupled to an underlying ground plane, as in Fig. 2.19, the surface potential V_s of the charge layer will be well defined. Specifically, if the layer has thickness d , the surface potential becomes

$$V_s = E \cdot d = \left(\frac{\rho_s}{\varepsilon}\right) d \tag{2.53}$$

The surface charge and its ground-plane image function as a double layer that introduces a potential jump between the ground plane and the upper surface of the insulator. The potential of a noncontacting voltmeter probe placed near the surface will be raised to the same potential V_s, allowing the surface charge ρ_s to be determined from Eq. (2.52).

If the charge on the insulator is not tightly coupled to a dominant ground plane, its surface potential will be strongly influenced by the position of the probe as well as by the insulator's position relative to other conductors and dielectrics. Under these conditions, the reading of the noncontacting voltmeter becomes extremely sensitive to probe position and cannot be determined without a detailed analysis of the fields surrounding the charge [45]. Such an analysis must account for two superimposed components: the field E_Q produced by the measured charge with the probe grounded, and the field E_V created by the voltage of the probe with the surface charge absent. The voltmeter will raise the probe potential until a null-field condition with $E_Q + E_V = 0$ is reached. Determining the relationship between the resulting probe voltage and the unknown surface charge requires a detailed field solution that takes into account the probe shape, probe position, and

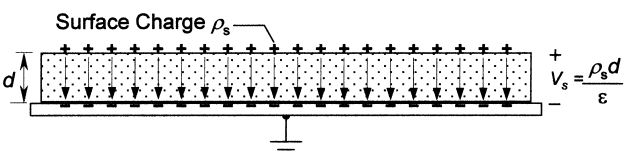


Figure 2.19 Surface charge on an insulator situated over a ground plane. The voltage on the surface of the insulator is clearly defined as $\rho_s d/\varepsilon$.

© 2006 by Taylor & Francis Group, LLC

insulator geometry. Because of the difficulty in translating voltmeter readings into actual charge values, noncontacting voltmeter measurements of isolated charge distributions that are not tightly coupled to ground planes are best used for relative measurement purposes only. A noncontacting voltmeter used in this way becomes particularly useful when measuring the decay time of a charge distribution. The position of the probe relative to the surface must remain fixed during such a measurement.

2.20. MICROMACHINES

The domain of micro-electromechanical systems, or MEMS, involves tiny microscale machines made from silicon, titanium, aluminum, or other materials. MEMS devices are fabricated using the tools of integrated-circuit manufacturing, including photolithography, pattern masking, deposition, and etching. Design solutions involving MEMS are found in many areas of technology. Examples include the accelerometers that deploy safety airbags in automobiles, pressure transducers, microfluidic valves, optical processing systems, and projection display devices.

One technique for making MEMS devices is known as *bulk micromachining*. In this method, microstructures are fabricated within a silicon wafer by a series of selective etching steps. Another common fabrication technique is called *surface micromachining*. The types of steps involved in the process are depicted in Fig. 2.20. A silicon substrate is patterned with alternating layers of polysilicon and oxide thin films that are used to build up the desired structure. The oxide films serve as *sacrificial layers* that support the

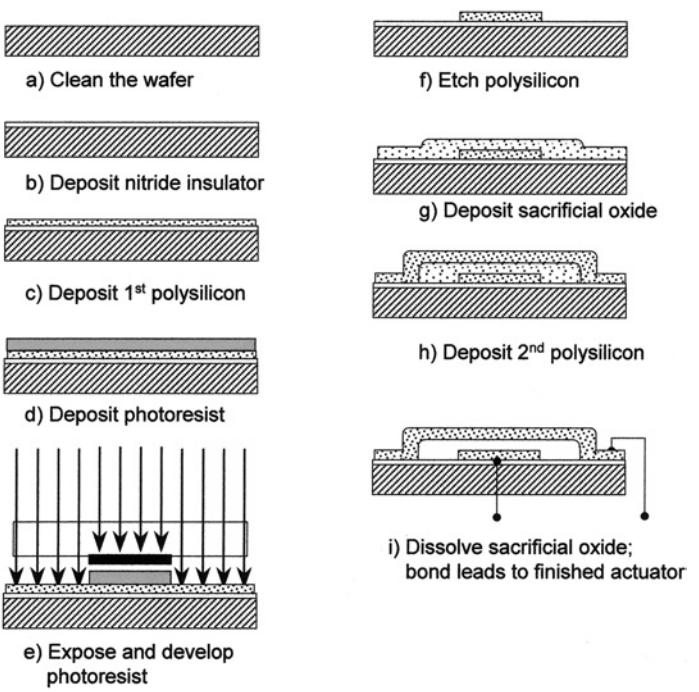

a) Clean the wafer
b) Deposit nitride insulator
c) Deposit 1st polysilicon
d) Deposit photoresist
e) Expose and develop photoresist
f) Etch polysilicon
g) Deposit sacrificial oxide
h) Deposit 2nd polysilicon
i) Dissolve sacrificial oxide; bond leads to finished actuator

Figure 2.20 Typical surface micromachining steps involved in MEMS fabrication. Oxides are used as sacrificial layers to produce structural members. A simple actuator is shown here.

© 2006 by Taylor & Francis Group, LLC

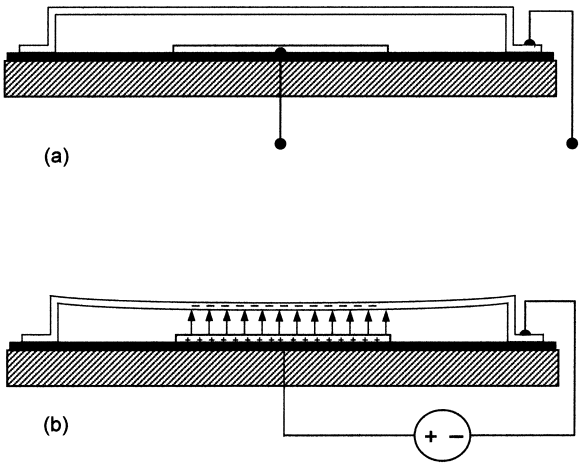


Figure 2.21 Applying a voltage to the actuator causes the membrane structure to deflect toward the substrate. The drawing is not to scale; typical width-to-gap spacing ratios are on the order of 100.

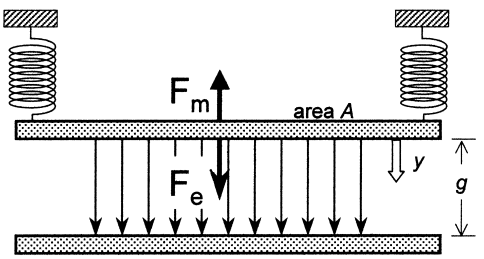


Figure 2.22 The MEMS actuator of Fig. 2.21 can be modeled by the simple mass-spring structure shown here. F_e is the electrostatic force when a voltage is applied; F_m is the mechanical restoring force.

polysilicon during sequential deposition steps but are removed in the final steps of fabrication. This construction technique is analogous to the way that stone arches were made in ancient times. Sand was used to support stone pieces and was removed when the building could support itself, leaving the finished structure.

One simple MEMS device used in numerous applications is illustrated in Fig. 2.21. This *double-cantilevered actuator* consists of a bridge supported over a fixed activation electrode. The bridge has a rectangular shape when viewed from the top and an aspect ratio (ratio of width to gap spacing) on the order of 100. When a voltage is applied between the bridge and the substrate, the electrostatic force of attraction causes the bridge to deflect downward. This vertical motion can be used to open and close valves, change the direction of reflected light, pump fluids, or mix chemicals in small micromixing chambers.

The typical bridge actuator has a gap spacing of a few microns and lateral dimensions on the order of 100 to 300 μm. This large aspect ratio allows the actuator to be modeled by the simple two-electrode capacitive structure shown in Fig. 2.22.

© 2006 by Taylor & Francis Group, LLC

The electrostatic force in the y direction can be found by taking the derivative of the stored energy (see Sec. 2.10):

$$F_E = \frac{\partial}{\partial y}\frac{1}{2}CV^2 = \frac{\varepsilon_0 A V^2}{(g-y)^2} \tag{2.54}$$

Here y is the deflection of the bridge, A its surface area, and g the gap spacing at zero deflection. As Eq. (2.54) shows, the electrostatic force increases with increasing deflection and becomes infinite as the residual gap spacing $(g-y)$ approaches zero. To first order, the mechanical restoring force will be proportional to the bridge deflection and can be expressed by the simple equation

$$F_M = -ky \tag{2.55}$$

The equilibrium deflection y for a given applied voltage will occur when $F_M = F_E$, i.e., when

$$ky = \frac{\varepsilon_0 A V^2}{(g-y)^2} \tag{2.56}$$

Figure 2.23 shows a plot of y versus V obtained from Eq. (2.56). For voltages above the critical value V_c, the mechanical restoring force can no longer hold back the electrostatic force, and the bridge collapses all the way to the underlying electrode. This phenomenon, known as *snap-through*, occurs at a deflection of one third of the zero-voltage gap spacing. It is reversible only by setting the applied voltage to zero and sometimes cannot be undone at all due to a surface adhesion phenomenon known as *sticktion*.

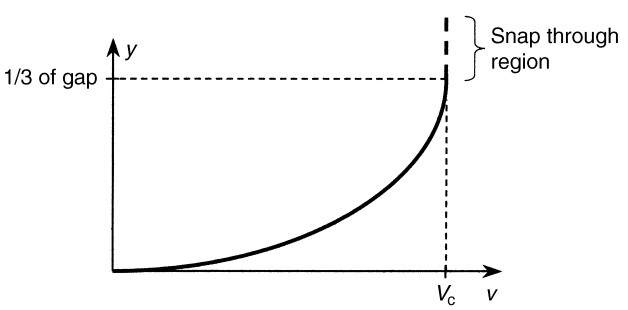


Figure 2.23 Voltage displacement curve for the actuator model of Fig. 2.22. At a deflection equal to one-third of the gap spacing, the electrostatic force overcomes the mechanical restoring force, causing the membrane to "snap through" to the substrate.

© 2006 by Taylor & Francis Group, LLC

The deflection at which snap-through occurs is easily derived by noting that at $v = V_c$, the slope of the voltage–displacement curve becomes infinite, i.e., dV/dy becomes zero. Equation (2.56) can be expressed in the form

$$V = \sqrt{\frac{ky}{\varepsilon_0 A}}(g - y) \tag{2.57}$$

The y derivative of this equation becomes zero when $y = g/3$.

2.21. DIGITAL MIRROR DEVICE

One interesting application of the MEMS actuator can be found in the digital mirror device (DMD) used in computer projection display systems. The DMD is an array of electrostatically-actuated micromirrors of the type shown in Fig. 2.24. Each actuator is capable of being driven into one of two bi-stable positions. When voltage is applied to the right-hand pad, as in Fig 2.24a, the actuator is bent to the right until it reaches its mechanical limit. Alternatively, when voltage is applied to the left-hand pad, as in Fig. 2.24b, the actuator bends to the left. The two deflection limits represent the logic **0** (no light projected) and logic **1** (light projected) states of the mirror pixel.

2.22. ELECTROSTATIC DISCHARGE AND CHARGE NEUTRALIZATION

Although much of electrostatics involves harnessing the forces of charge, sometimes static electricity can be most undesirable. Unwanted electrostatic forces can interfere with materials and devices, and sparks from accumulated charge can be quite hazardous in the vicinity of flammable liquids, gases, and air dust mixtures [12, 46–51]. In this section, we examine situations in which electrostatics is a problem and where the main objective is to eliminate its effects.

Many manufacturing processes involve large moving webs of insulating materials, such as photographic films, textiles, food packaging materials, and adhesive tapes. These materials can be adversely affected by the presence of static electricity. A moving web is easily charged by contact electrification because it inevitably makes contact with rollers, guide plates, and other processing structures. These contact and separation events provide ample opportunity for charge separation to occur [52]. A charged web can be attracted to parts of the processing machinery, causing jams in the machinery or breakage of the web material. In some situations, local surface sparks may also occur that can ruin the

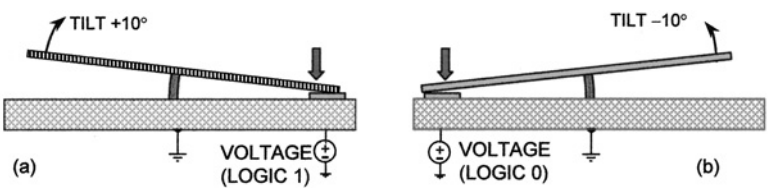


Figure 2.24 Simplified schematic of digital mirror device. Each pixel tilts $\pm 10^\circ$ in response to applied voltages.

© 2006 by Taylor & Francis Group, LLC

processed material. This issue is especially important in the manufacturing of photographic films, which can be prematurely exposed by the light from sparks or other discharges.

Electrostatic charge is very undesirable in the semiconductor industry. Sensitive semiconductor components, particularly those that rely on metal-oxide-semiconductor (MOS) technology, can be permanently damaged by the electric fields from nearby charged materials or by the discharges that occur when charged materials come into contact with grounded conductors. Discharges similar to the ''carpet sparks'' that plague temperate climates in winter can render semiconductor chips useless. A static-charged wafer also can attract damaging dust particles and other contaminants.

The term *electrostatic discharge* (ESD) refers to any unwanted sparking event caused by accumulated static charge. An abundance of books and other resources may be found in the literature to aid the electrostatics professional responsible for preventing ESD in a production facility [53–58].

Numerous methods exist to neutralize accumulated charge before it can lead to an ESD event. The ionizing neutralizer is one of the more important devices used to prevent the build up of unwanted static charge. An ionizer produces both positively and negatively charged ions of air that are dispersed in proximity to sensitive devices and work areas. When undesirable charge appears on an object from contact electrification or induction charging, ions of the opposite polarity produced by the ionizer are attracted to the object and quickly neutralize it. The relatively high mobility of air ions allows this neutralization to occur rapidly, usually in a matter of seconds or less.

The typical ionizer produces ions via the process of corona discharge. A coronating conductor, usually a sharp needle point, or sometimes a thin, axially mounted wire, is energized to a voltage on the order of 5 to 10 kV. An extremely high electric field develops at the electrode, causing electrons to be stripped from neutral air molecules via an avalanche multiplication process (see Sec. 2.9). In order to accommodate unwanted charge of either polarity, and to avoid inadvertent charging of surfaces, the ionizer must simultaneously produce balanced quantities of positive and negative charge. Some ionizers produce bipolar charge by applying an ac voltage to the corona electrode. The ionizer thus alternately produces positive and negative ions that migrate as a bipolar charge cloud toward the work piece. Ions having polarity opposite the charge being neutralized will be attracted to the work surface, while ions of the same polarity will be repelled. The undesired charge thus extracts from the ionizer only what it needs to be neutralized.

Other ionizers use a different technique in which adjacent pairs of electrodes are energized simultaneously, one with positive and the other with negative dc high voltage. Still other neutralizers use separate positive and negative electrodes, but energize first the positive side, then the negative side for different intervals of time. Because positive and negative electrodes typically produce ions at different rates, this latter method of electrification allows the designer to adjust the ''on'' times of each polarity, thereby ensuring that the neutralizer produces the proper balance of positive and negative ions.

Although the production of yet more charge may seem a paradoxical way to eliminate unwanted charge, the key to the method lies in maintaining a proper balance of positive and negative ions produced by the ionizer, so that no additional net charge is imparted to nearby objects or surfaces. Thus, one figure of merit for a good ionizer is its overall balance as measured by the lack of charge accumulation of either polarity at the work piece served by the ionizer. Another figure of merit is the speed with which an ionizer can neutralize unwanted charge. This parameter is sometimes called the ionizer's *effectiveness*. The more rapidly unwanted static charge can be neutralized, the less

© 2006 by Taylor & Francis Group, LLC

chance it will have to affect sensitive electronic components or interfere with a production process. Effectiveness of an ionizer is maximized by transporting the needed charge as rapidly as possible to the neutralized object [21]. Sometimes this process is assisted by air flow from a fan or blowing air stream. Increasing the density of ions beyond some minimum level does not increase effectiveness because the extra ions recombine quickly.

2.23. SUMMARY

This chapter is intended to serve as an introduction to the many applications of electrostatics in science, technology, and industry. The topics presented are not all inclusive of this fascinating and extensive discipline, and the reader is encouraged to explore some of the many reference books cited in the text. Despite its long history [59], electrostatics is an ever-evolving field that seems to emerge anew with each new vista of discovery.

REFERENCES

1. Schein, L.B.; LaHa, M.; Novotny, D. Theory of insulator charging. Phys. Lett. **1992**, *A 167*, 79–83.
2. Horn, R.G.; Smith, D.T. Contact electrification and adhesion between dissimilar materials. Science **1992**, *256*, 362–364.
3. Harper, W.R. Contact and frictional electrification. In *Monographs on the Physics and Chemistry of Materials*; Clarendon Press: Oxford, 1967.
4. Shinbrot, T. A look at charging mechanisms. J. Electrostat. **1985**, *17*, 113–123.
5. Davies, D.K. Charge generation of dielectric surfaces. J. Phys. **1969**, *D2*, 1533.
6. Schein, L.B.; Cranch, J. The static electrification of mixtures of insulating powders. J. Appl. Phys. **1975**, *46*, 5140.
7. Schein, L.B.; Castle, G.S.P.; Dean, A. Theory of monocomponent development. J. Imag. Technol **1989**, *15*, 9.
8. Schein, L.B.; LaHa, M.; Novotny, D. Theory of insulator charging. Phys. Lett. **1992**, *A 167*, 79.
9. Cross, J. *Electrostatics: Principles, Problems and Applications*; IOP Publishing: Bristol, 1987; 500.
10. Taylor, D.M.; Secker, P.E. *Industrial Electrostatics*; John Wiley and Sons: New York, 1994.
11. Montgomery, D.J. Static electrification in solids. Solid State Phys. **1959**, *9*, 139–197.
12. Glor, M. *Electrostatic Hazards in Powder Handling*; John Wiley and Sons: New York, 1988.
13. Coehn, A. Ann. Physik, **1898**, *64*, 217.
14. JW (Lord) Raleigh, On the equilibrium of liquid conducting masses charged with electricity. Phil. Mag. **1882**, *14*, 184–186.
15. Melcher, J.R. *Continuum Electromechanics*; MIT Press: Cambridge, Massachusetts, 1981, 8.44.
16. Bailey, A.G. *Electrostatic Spraying of Liquids*; John Wiley and Sons: New York, 1988.
17. Law, S.E. Electrostatic atomization and spraying. In *Handbook of Electrostatic Processes*; Chang, J.S., Kelly, A.J., Crowley, J.M., Eds.; Marcel Dekker: New York, 1995; 413–440.
18. Cobine, J.D. *Gaseous Conductors*; Dover Press: New York, 1958, 252–281.
19. Tobazéon, R. Electrical phenomena of dielectric materials. In *Handbook of Electrostatic Processes*; Chang, J.S., Kelly, A.J., Crowley, J.M., Eds.; Marcel Dekker: New York, 1995; 51–82.
20. Peek, F.W. *Dielectric Phenomena in High Voltage Engineering*; McGraw-Hill: New York, 1929, 48–108.
21. Crowley, J.M. *Fundamentals of Applied Electrostatics*; Wiley: New York, 1986, 164, 207–225.

© 2006 by Taylor & Francis Group, LLC

22. Haus, H.; Melcher, J.R. *Electromagnetic Fields and Energy*; Prentice-Hall: Englewood Cliffs, NJ, 1989, 486–521.
23. Woodson, H.; Melcher, J.R. *Electromechanical Dynamics*; John Wiley and Sons: New York, 1968, Chapter 8.
24. Zahn, M., *Electromagnetic Field Theory: A Problem Solving Approach*; John Wiley and Sons: New York, 1979, 204–230.
25. Law, S.E. Electrostatic pesticide spraying: concepts and practice. IEEE Trans. **1983**, *IA-19* (2), 160–168.
26. Inculet, I.I.; Fisher, J.K. Electrostatic aerial spraying. IEEE Trans. **1989**, *25* (3).
27. Pauthenier, M.M.; Moreau-Hanot, M. La charge des particules spheriques dans un champ ionize. J. Phys. Radium (Paris) **1932**, *3*, 590.
28. Schein, L.B. *Electrophotography and Development Physics*; 2nd Ed.; Springer Verlag: New York, 1992.
29. White, H.J. *Industrial Electrostatic Precipitation*; Reading, Addison-Wesley: MA, 1962.
30. Masuda, S.; Hosokawa, H. *Electrostatic precipitation*. In *Handbook of Electrostatics*; Chang, J.S., Kelly, A.J., Crowley, J.M., Eds.; Marcel Dekker: New York, 1995; 441–480.
31. Masuda, S. Electrical precipitation of aerosols. Proc. 2nd Aerosol Int. Conf., Berlin, Germany: Pergamon Press, 1986; 694–703.
32. White, H.J. Particle charging in electrostatic precipitation. AIEE Trans. Pt. 1, *70*, 1186.
33. Masuda, S.; Nonogaki, Y. Detection of back discharge in electrostatic precipitators. Rec. IEEE/IAS Annual Conference, Cincinnati, Ohio, 1980; 912–917.
34. Masuda, S.; Obata, T.; Hirai, J. A pulse voltage source for electrostatic precipitators. Rec. IEEE/IAS Conf., Toronto, Canada, 1980; 23–30.
35. Nyberg, B.R.; Herstad, K.; Larsen, K.B.; Hansen, T. Measuring electric fields by using pressure sensitive elements. IEEE Trans. Elec. Ins, **1979**, *EI-14*, 250–255.
36. Horenstein, M. A direct gate field-effect transistor for the measurement of dc electric fields. IEEE Trans. Electr. Dev. **1985**, *ED-32* (3): 716.
37. McCaslin, J.B. Electrometer for ionization chambers using metal-oxide-semiconductor field-effect transistors. Rev. Sci. Instr. **1964**, *35* (11), 1587.
38. Blitshteyn, M. Measuring the electric field of flat surfaces with electrostatic field meters. Evaluation Engineering, **Nov. 1984**, *23* (10), 70–86.
39. Schwab, A.J. *High Voltage Measurement Techniques*; MIT Press: Cambridge, MA, 1972, 97–101.
40. Secker, P.E. Instruments for electrostatic measurements. J. Elelectrostat. **1984**, *16* (1), 1–19.
41. Vosteen, R.E.; Bartnikas, R. Electrostatic charge measurement. Engnr Dielectrics, Vol IIB, Electr Prop Sol Insul Matls, ASTM Tech Publ 926, 440–489.
42. Vosteen, W. A high speed electrostatic voltmeter technique. Proc IEEE Ind Appl Soc Annual Meeting IAS-88(2): 1988; 1617–1619.
43. Horenstein, M. Measurement of electrostatic fields, voltages, and charges. In *Handbook of Electrostatics*; Chang, J.S., Kelly, A.J., Crowley, J.M. Eds.; Marcel Dekker: New York, 1995; 225–246.
44. Popovic, Z.; Popovic, B.D. *Introductory Electromagnetics*; Prentice-Hall: Upper Saddle River, NJ, 2000; 114–115.
45. Horenstein, M. Measuring surface charge with a noncontacting voltmeter. J. Electrostat. **1995**, *35*, 2.
46. Gibson, N.; Lloyd, F.C. Incendivity of discharges from electrostatically charged plastics. Brit. J. Appl. Phys. **1965**, *16*, 619–1631.
47. Gibson, N. Electrostatic hazards. In Electrostatics '83; Inst. Phys. Conf. Ser. No. 66, Oxford, 1983; 1–11.
48. Glor, M. Hazards due to electrostatic charging of powders. J. Electrostatics **1985**, *16*, 175–181.
49. Pratt, T.H. *Electrostatic Ignitions of Fires and Explosions*; Burgoyne: Marietta, GA, 1997, 115–152.

© 2006 by Taylor & Francis Group, LLC

50. Lüttgens, G.; Wilson, N. *Electrostatic Hazards*; Butterworth-Heinemann: Oxford, 1997, 137–148.
51. Bailey, A.G. Electrostatic hazards during liquid transport and spraying. In *Handbook of Electrostatics*; Chang, J.S., Kelly, A.J., Crowley, J.M., Eds.; Marcel Dekker: New York, 1995; 703–732.
52. Hughes, J.F.; Au, A.M.K.; Blythe, A.R. Electrical charging and discharging between films and metal rollers. Electrostatics '79. Inst. Phys. Conf. Ser. No. 48, Oxford, 1979; 37–44.
53. Horvath, T.; Berta, I. *Static Elimination*; Research Studies Press: New York, 1982; 118.
54. Davies, D.K. Harmful effects and damage to electronics by electrostatic discharges. J. Electrostatics **1985**, *16*, 329–342.
55. McAteer, O.J.; Twist, R.E. Latent ESD failures, EOS/ESD Symposium Proceedings, Orlando, FL, 1982; 41–48.
56. Boxleitner, W. *Electrostatic Discharge and Electronic Equipment: A Practical Guide for Designing to Prevent ESD Problems*; IEEE Press: New York, 1989, 73–84.
57. McAteer, O.J. *Electrostatic Discharge Control*; McGraw-Hill: New York, 1990.
58. Greason, W. *Electrostatic Discharge in Electronics*; John Wiley and Sons: New York, 1993.
59. Moore, A.D. *Electrostatics and Its Applications*; John Wiley and Sons: New York, 1973.

© 2006 by Taylor & Francis Group, LLC

3
Magnetostatics

Milica Popović
McGill University
Montréal, Quebec

Branko D. Popović[†]
University of Belgrade
Belgrade, Yugoslavia

Zoya Popović
University of Colorado
Boulder, Colorado

To the loving memory of our father, professor, and coauthor. We hope that he would have agreed with the changes we have made after his last edits.
— *Milica and Zoya Popović*

3.1. INTRODUCTION

The force between two static electric charges is given by Coulomb's law, obtained directly from measurements. Although small, this force is easily measurable. If two charges are *moving*, there is an *additional* force between them, the *magnetic force*. The magnetic force between *individual* moving charges is extremely small when compared with the Coulomb force. Actually, it is so small that it cannot be detected experimentally between just a pair of moving charges. However, these forces can be measured using a vast number of electrons (practically one per atom) in organized motion, i.e., electric current. Electric current exists within almost electrically neutral materials. Thus, magnetic force can be measured independent of electric forces, which are a result of charge unbalance.

Experiments indicate that, because of this vast number of interacting moving charges, the magnetic force between two current-carrying conductors can be much larger than the maximum electric force between them. For example, strong electromagnets can carry weights of several tons, while electric force cannot have even a fraction of that strength. Consequently, the magnetic force has many applications. For example, the approximate direction of the North Magnetic Pole is detected with a magnetic device—a compass. Recording and storing various data are most commonly accomplished using

[†]Deceased.

© 2006 by Taylor & Francis Group, LLC

magnetic storage components, such as computer disks and tapes. Most household appliances, as well as industrial plants, use motors and generators, the operation of which is based on magnetic forces.

The goal of this chapter is to present:

Fundamental theoretical foundations for magnetostatics, most importantly Ampere's law
Some simple and commonly encountered examples, such as calculation of the magnetic field inside a coaxial cable
A few common applications, such as Hall element sensors, magnetic storage, and MRI medical imaging.

3.2. THEORETICAL BACKGROUND AND FUNDAMENTAL EQUATIONS

3.2.1. Magnetic Flux Density and Lorentz Force

The electric force on a charge is described in terms of the electric field vector, **E**. The magnetic force on a charge moving with respect to other moving charges is described in terms of the *magnetic flux density vector*, **B**. The unit for **B** is a *tesla* (T). If a point charge Q [in coulombs (C)] is moving with a velocity **v** [in meters per second (m/s)], it experiences a force [in newtons (N)] equal to

$$\mathbf{F} = Q\mathbf{v} \times \mathbf{B} \tag{3.1}$$

where "×" denotes the vector product (or cross product) of two vectors.

The region of space in which a force of the form in Eq. (3.1) acts on a moving charge is said to have a *magnetic field* present. If in addition there is an electric field in that region, the total force on the charge (the Lorentz force) is given by

$$\mathbf{F} = Q\mathbf{E} + Q\mathbf{v} \times \mathbf{B} \tag{3.2}$$

where **E** is the electric field intensity in volts per meter (V/m).

3.2.2. The Biot–Savart Law

The magnetic flux density is produced by current-carrying conductors or by permanent magnets. If the source of the magnetic field is the electric current in thin wire loops, i.e. current loops, situated in vacuum (or in air), we first adopt the orientation along the loop to be in the direction of the current in it. Next we define the product of the wire current, I, with a short vector length of the wire, $d\mathbf{l}$ (in the adopted reference direction along the wire), as the *current element*, $I\,d\mathbf{l}$ (Fig. 3.1a). With these definitions, the magnetic flux density due to the entire current loop C (which may be arbitrarily complex), is at any point given by the experimentally obtained Biot–Savart law:

$$\mathbf{B} = \frac{\mu_0}{4\pi}\oint_C \frac{I\,d\mathbf{l} \times \mathbf{a}_r}{r^2} \tag{3.3}$$

The unit vector $\mathbf{a}_r$ is directed *from the source point* (i.e., the current element) *towards the field point* (i.e., the point at which we determine **B**). The constant μ_0 is known as the

© 2006 by Taylor & Francis Group, LLC

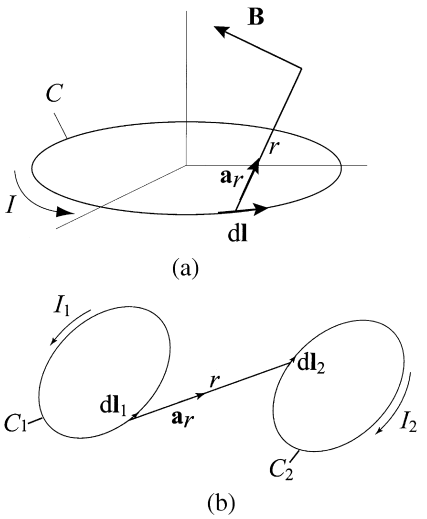


Figure 3.1 (a) A current loop with a current element. (b) Two current loops and a pair of current elements along them.

permeability of vacuum. Its value is defined to be exactly

$$\mu_0 = 4\pi \times 10^{-7}\ \mathrm{H/m}$$

Note that the magnetic flux density vector of individual current elements is perpendicular to the plane of the vectors $\mathbf{r}$ and $d\mathbf{l}$. Its orientation is determined by the right-hand rule when the vector $d\mathbf{l}$ is rotated by the shortest route towards the vector $\mathbf{a}_r$. The (vector) integral in Eq. (3.3) can be evaluated in closed form in only a few cases, but it can be always evaluated numerically.

The line current I in Eq. (3.3) is an approximation to volume current. Volume currents are described by the current density vector, $\mathbf{J}$ [amperes per meter squared ($\mathrm{A/m^2}$)]. Let ΔS be the cross-sectional area of the wire. The integral in Eq. (3.3) then becomes a volume integral where $I\,d\mathbf{l}$ is replaced by $\mathbf{J} \cdot \Delta S \cdot d\mathbf{l} = \mathbf{J} \cdot dv$. At high frequencies (above about 1MHz), currents in metallic conductors are distributed in very thin layers on conductor surfaces (the *skin effect*). These layers are so thin that they can be regarded as geometrical surfaces. In such cases we introduce the concept of *surface current density* $\mathbf{J}_s$ (in A/m), and the integral in Eq. (3.3) becomes a surface integral, where $I\,d\mathbf{l}$ is replaced by $\mathrm{J}_s\,dS$.

3.2.3. Units: How Large is a Tesla?

The unit for magnetic flux density in the SI system is a tesla* (T). A feeling for the magnitude of a tesla can be obtained from the following examples. The magnetic flux density of the earth's dc magnetic field is on the order of 10^{-4} T. Around current-carrying

*The unit was named after the American scientist of Serbian origin Nikola Tesla, who won the ac–dc battle over Thomas Edison and invented three-phase systems, induction motors, and radio transmission. An excellent biography of this eccentric genius is *Tesla, Man out of Time*, by Margaret Cheney, Dorset Press, NY, 1989.

© 2006 by Taylor & Francis Group, LLC

conductors in vacuum, the intensity of **B** ranges from about 10^{-6} T to about 10^{-2} T. In air gaps of electrical machines, the magnetic flux density can be on the order of 1 T. Electromagnets used in magnetic-resonance imaging (MRI) range from about 2 T to about 4 T [5,15]. Superconducting magnets can produce flux densities of several dozen T.

3.2.4. Magnetic Force

From Eq. (3.2) it follows that the magnetic force on a current element $I\,d\mathbf{l}$ in a magnetic field of flux density **B** is given by

$$d\mathbf{F} = I\,d\mathbf{l} \times \mathbf{B} \quad \text{(N)} \tag{3.4}$$

Combining this expression with the Biot–Savart law, an expression for the magnetic force between two current loops C_1 and C_2 (Fig. 3.1b) is obtained:

$$d\mathbf{F}_{C1 \text{ on } C2} = \frac{\mu_0}{4\pi} \oint_{C_1} \oint_{C_2} I_2\,d\mathbf{l}_2 \times I_1 d\mathbf{l}_1 \tag{3.5}$$

3.2.5. Magnetic Moment

For a current loop of vector area **S** (the unit vector normal to **S**, by convention, is determined by the right-hand rule with respect to the reference direction along the loop), the *magnetic moment* of the loop, **m**, is defined as

$$\mathbf{m} = I \times \mathbf{S} \tag{3.6}$$

If this loop is situated in a uniform magnetic field of magnetic flux density **B**, the *mechanical* moment, **T**, on the loop resulting from magnetic forces on its elements is

$$\mathbf{T} = \mathbf{m} \times \mathbf{B} \tag{3.7}$$

This expression is important for understanding applications such as motors and generators.

The lines of vector **B** are defined as (generally curved) imaginary lines such that vector **B** is tangential to them at all points. For example, from Eq. (3.3) it is evident that the lines of vector **B** for a single current element are circles centered along the line of the current element and in planes perpendicular to the element.

3.2.6. Magnetic Flux

The flux of vector **B** through a surface is termed the *magnetic flux*. It plays a very important role in magnetic circuits, and a fundamental role in one of the most important electromagnetic phenomena, electromagnetic induction. The magnetic flux, Φ, through a surface S is given by

$$\Phi = \int_S \mathbf{B} \cdot d\mathbf{S} \qquad \text{in webers (Wb)} \tag{3.8}$$

© 2006 by Taylor & Francis Group, LLC

The magnetic flux has a very simple and important property: it is equal to zero through *any* closed surface,

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = 0 \tag{3.9}$$

This relation is known as the *law of conservation of magnetic flux* and represents the fourth Maxwell's equation in integral form. In differential form, it can be written as $\nabla \cdot \mathbf{B} = 0$, using the divergence theorem. An interpretation of the law of conservation of magnetic flux is that "magnetic charges" do not exist, i.e., a south and north pole of a magnet are never found separately. The law tells us also that the lines of vector **B** do not have a beginning or an end. Sometimes, this last statement is phrased more loosely: it is said that the lines of vector **B** close on themselves.

An important conclusion follows: If we have a closed contour C in the field and imagine any number of surfaces spanned over it, *the magnetic flux through any such surface*, *spanned over the same contour*, *is the same*. There is just one condition that needs to be fulfilled in order for this to be true: the unit vector normal to all the surfaces must be the same with respect to the contour, as shown in Fig. 3.2. It is customary to orient the contour and then to define the vector unit normal on any surface on it according to the right-hand rule.

3.2.7. Ampere's Law in Vacuum

The magnetic flux density vector **B** resulting from a time-invariant current density **J** has a very simple and important property: If we compute the line integral of **B** along any closed contour C, it will be equal to μ_0 times the total current that flows through any surface spanned over the contour. This is *Ampere's law* for dc (time-invariant) currents in vacuum (Fig. 3.3):

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S} \tag{3.10}$$

The reference direction of the vector surface elements of S is adopted according to the right-hand rule with respect to the reference direction of the contour. In the applications of Ampere's law, it is very useful to keep in mind that the flux of the current density vector (the current intensity) is the same through all surfaces having a common boundary

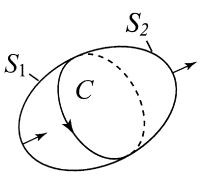


Figure 3.2 Two surfaces, S_1 and S_2, defined by a common contour C, form a closed surface to which the law of conservation of magnetic flux applies—the magnetic flux through them is the same. The direction chosen for the loop determines the normal vector directions for S_1 and S_2 according to the right-hand rule.

© 2006 by Taylor & Francis Group, LLC

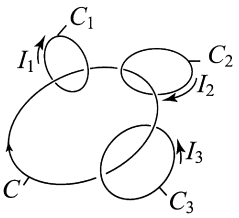


Figure 3.3 Three current loops and the illustration of Ampere's law. The line integral of **B** along C in the case shown equals I_1–I_2–I_3.

contour. Ampere's law is not a new property of the magnetic field—it follows from the Biot–Savart law, which in turn is based on experiment.

Ampere's law in Eq. (3.10) is a general law of the magnetic field of time-invariant (dc) currents in vacuum. It can be extended to cases of materials in the magnetic field, but in this form it is *not* valid for magnetic fields produced by *time-varying* (ac) currents. Since the left-hand side in Ampere's law is a vector integral, while the right-hand side is a scalar, it can be used to determine analytically vector **B** only for problems with a high level of symmetry for which the vector integral can be reduced to a scalar one. Several such practical commonly encountered cases are a cylindrical wire, a coaxial cable and parallel flat current sheets.

3.2.8. Magnetic Field in Materials

If a body is placed in a magnetic field, magnetic forces act on all *moving* charges within the atoms of the material. These moving charges make the atoms and molecules inside the material look like tiny current loops. The moment of magnetic forces on a current loop, Eq. (3.7), tends to align vectors **m** and **B**. Therefore, in the presence of the field, a substance becomes a large aggregate of oriented elementary current loops which produce their own magnetic fields. Since the rest of the body does not produce any magnetic field, a substance in the magnetic field can be visualized as a large set of oriented elementary current loops situated in vacuum. A material in which magnetic forces produce such oriented elementary current loops is referred to as a *magnetized* material. It is possible to replace a material body in a magnetic field with equivalent *macroscopic* currents *in vacuum* and analyze the magnetic field provided that we know how to find these equivalent currents. Here the word *macroscopic* refers to the fact that a small volume of a material is assumed to have a very large number of atoms or molecules.

The number of revolutions per second of an electron around the nucleus is very large—about 10^{15} revolutions/s. Therefore, it is reasonable to say that such a rapidly revolving electron is a small (elementary) current loop with an associated magnetic moment. This picture is, in fact, more complicated since in addition electrons have a magnetic moment of their own (their spin). However, each atom can macroscopically be viewed as an equivalent single current loop. Such an elementary current loop is called an *Ampere current*. It is characterized by its magnetic moment, $\mathbf{m} = I\mathbf{S}$. The macroscopic quantity called the *magnetization vector*, **M**, describes the density of the vector magnetic moments in a magnetic material at a given point and for a substance with N Ampere currents per unit volume can be written as

$$\mathbf{M} = \sum \frac{\mathbf{m}_{\text{in } dv}}{dv} = N\mathbf{m} \tag{3.11}$$

© 2006 by Taylor & Francis Group, LLC

The significance of Eq. (3.11) is as follows. The magnetic field of a single current loop in vacuum can be determined from the Biot–Savart law. The vector **B** of such a loop at large distances from the loop (when compared with the loop size) is proportional to the magnetic moment, **m**, of the loop. According to Eq. (3.11) we can subdivide magnetized materials into small volumes, ΔV, each containing a very large number of Ampere currents, and represent each volume by a single larger Ampere current of moment $\mathbf{M}\,\Delta V$. Consequently, if we determine the magnetization vector at all points, we can find vector **B** by integrating the field of these larger Ampere currents over the magnetized material. This is much simpler than adding the fields of the prohibitively large number of individual Ampere currents.

3.2.9. Generalized Ampere's Law and Magnetic Field Intensity

Ampere's law in the form as in Eq. (3.10) is valid for any current distribution *in vacuum*. Since the magnetized substance is but a vast number of elementary current loops in vacuum, we can apply Ampere's law to fields in materials, provided we find how to include these elementary currents on the right-hand side of Eq. (3.10). The total current of elementary current loops "strung" along a closed contour C, i.e., the total current of all Ampere's currents through the surface S defined by contour C, is given by

$$I_{\text{Ampere through } S} = \oint_C \mathbf{M} \cdot d\mathbf{l} \tag{3.12}$$

The generalized form of Ampere's law valid for time-invariant currents therefore reads

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 \left(\int_S \mathbf{J} \cdot d\mathbf{S} + \oint_C \mathbf{M} \cdot d\mathbf{l} \right) \tag{3.13}$$

Since the contour C is the same for the integrals on the left-hand and right-hand sides of the equation, this can be written as

$$\oint_C \left(\frac{\mathbf{B}}{\mu_0} - \mathbf{M} \right) \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S} \tag{3.14}$$

The combined vector $\mathbf{B}/\mu_0 - \mathbf{M}$ has a convenient property: Its line integral along any closed contour depends only on the *actual* current through the contour. This is the only current we can control—switch it on and off, change its intensity or direction, etc. Therefore, the combined vector is defined as a new vector that describes the magnetic field in the presence of materials, known as the *magnetic field intensity*, **H**:

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M} \qquad (\text{A/m}) \tag{3.15}$$

With this definition, the generalized Ampere's law takes the final form:

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S} \tag{3.16}$$

© 2006 by Taylor & Francis Group, LLC

As its special form, valid for currents in vacuum, this form of Ampere's law is also valid *only for time-constant* (*dc*) *currents*.

The definition of the magnetic field intensity vector in Eq. (3.15) is general and valid for any material. Most materials are those for which the magnetization vector, **M**, is a linear function of the local vector **B** (the cause of material magnetization). In such cases a linear relationship exists between any two of the three vectors **H**, **B**, and **M**. Usually, vector **M** is expressed as

$$\mathbf{M} = \chi_m \mathbf{H} \qquad (\chi_m \text{ is dimensionless, } \mathbf{M} \text{ in A/m}) \tag{3.17}$$

The dimensionless factor χ_m is known as the *magnetic susceptibility* of the material. We then use Eq. (3.17) and express **B** in terms of **H**:

$$\mathbf{B} = \mu_0(1 + \chi_m)\mathbf{H} = \mu_0 \mu_r \mathbf{H} \qquad (\mu_r \text{ is dimensionless, } \mu_0 \text{ in H/m}) \tag{3.18}$$

The dimensionless factor $\mu_r = (1 + \chi_m)$ is known as the *relative permeability* of the material, and μ as the *permeability* of the material. Materials for which Eq. (3.18) holds are *linear* magnetic materials. If it does not hold, they are *nonlinear*. If at all points of the material μ is the same, the material is said to be *homogeneous*; otherwise, it is *inhomogeneous*.

Linear magnetic materials can be *diamagnetic*, for which $\chi_m < 0$ (i.e., $\mu_r < 1$), or *paramagnetic*, for which $\chi_m > 0$ (i.e., $\mu_r > 1$). For both diamagnetic and paramagnetic materials $\mu_r \cong 1$, differing from unity by less than ± 0.001. Therefore, in almost all applications diamagnetic and paramagnetic materials can be considered to have $\mu = \mu_0$.

Ampere's law in Eq. (3.16) can be transformed into a differential equation, i.e., its differential form, by applying Stokes' theorem of vector analysis:

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{3.19}$$

This differential form of the generalized Ampere's law is valid only for time-invariant currents and magnetic fields.

3.2.10. Macroscopic Currents Equivalent to a Magnetized Material

The macroscopic currents *in vacuum* equivalent to a magnetized material can be both volume and surface currents. The volume density of these currents is given by

$$\mathbf{J}_m = \nabla \times \mathbf{M} \qquad (\text{A/m}^2) \tag{3.20}$$

This has a practical implication as follows. In case of a linear, homogeneous material of magnetic susceptibility χ_m, with no macroscopic currents in it,

$$\mathbf{J}_m = \nabla \times \mathbf{M} = \nabla \times (\chi_m \mathbf{H}) = \chi_m \nabla \times \mathbf{H} = \mathbf{0} \tag{3.21}$$

since $\nabla \times \mathbf{H} = \mathbf{0}$ if $\mathbf{J} = \mathbf{0}$, as assumed. Consequently, in a linear and homogeneous magnetized material with no macroscopic currents there is no volume distribution of equivalent currents. This conclusion is relevant for determining magnetic fields of magnetized materials, where the entire material can be replaced by equivalent surface

© 2006 by Taylor & Francis Group, LLC

currents. For example, the problem of a magnetized cylinder reduces to solving the simple case of a solenoid (coil).

3.2.11. Boundary Conditions

Quite often it is necessary to solve magnetic problems involving inhomogeneous magnetic materials that include boundaries. To be able to do this it is necessary to know the relations that must be satisfied by various magnetic quantities at two close points on the two sides of a boundary surface. Such relations are called *boundary conditions*. The two most important boundary conditions are those for the tangential components of **H** and the normal components of **B**. Assuming that there are no macroscopic surface currents on the boundary surface, from the generalized form of Ampere's law it follows that the tangential components of **H** are equal:

$$\mathbf{H}_{1\text{tang}} = \mathbf{H}_{2\text{tang}} \tag{3.22}$$

The condition for the normal components of **B** follows from the law of conservation of magnetic flux, Eq. (3.8), and has the form

$$\mathbf{B}_{1\text{norm}} = \mathbf{B}_{2\text{norm}} \tag{3.23}$$

The boundary conditions in Eqs. (3.22) and (3.23) are valid for *any* media—linear or nonlinear. If the two media are linear, characterized by permeabilities μ_1 and μ_2, the two conditions can be also written in the form

$$\frac{\mathbf{B}_{1\text{tang}}}{\mu_1} = \frac{\mathbf{B}_{2\text{tang}}}{\mu_2} \tag{3.24}$$

and

$$\mu_1 \mathbf{H}_{1\text{norm}} = \mu_2 \mathbf{H}_{2\text{norm}} \tag{3.25}$$

If two media divided by a boundary surface are linear, the lines of vector **B** and **H** refract on the surface according to a simple rule, which follows from the boundary conditions. With reference to Fig. 3.4, this rule is of the form

$$\frac{\tan\alpha_1}{\tan\alpha_2} = \frac{\mu_1}{\mu_2} \tag{3.26}$$

On a boundary between two magnetized materials, Fig. 3.5, the equivalent surface current density is given by

$$\mathbf{J}_{\text{ms}} = \mathbf{n} \times (\mathbf{M}_1 - \mathbf{M}_2) \tag{3.27}$$

Note that the unit vector **n** normal to the boundary surface is directed into medium 1 (Fig. 3.5).

The most interesting practical case of refraction of magnetic field lines is on the boundary surface between air and a medium of high permeability. Let air be medium 1.

© 2006 by Taylor & Francis Group, LLC

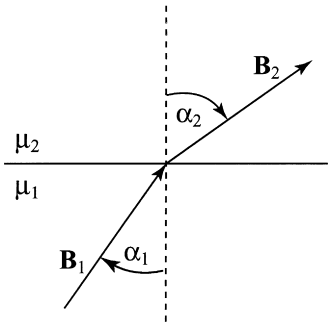


Figure 3.4 Lines of vector **B** or vector **H** refract according to Eq. (3.26).

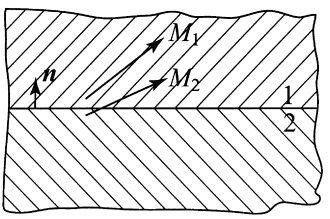


Figure 3.5 Boundary surface between two magnetized materials.

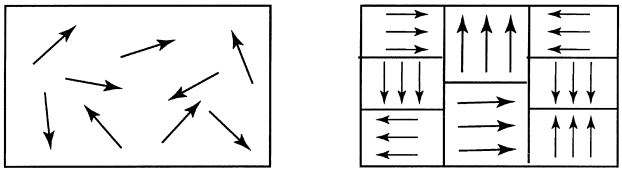
Then the right-hand side of Eq. (3.26) is very small. This means that $\tan\alpha_1$ must also be very small *for any* α_2 (except if $\alpha_2 = \pi/2$, i.e., if the magnetic field lines in the medium of high permeability are tangential to the boundary surface). Since for small angles $\tan\alpha_1 \cong \alpha_1$, the *magnetic field lines in air are practically normal to the surface of high permeability*. This conclusion is very important in the analysis of electrical machines with cores of high permeability, magnetic circuits (such as transformers), etc.

3.2.12. Basic Properties of Magnetic Materials

In the absence of an external magnetic field, atoms and molecules of many materials have no magnetic moment. Such materials are referred to as *diamagnetic materials*. When brought into a magnetic field, a current is induced in each atom and has the effect of reducing the field. (This effect is due to electromagnetic induction, and exists in *all* materials. It is very small in magnitude, and in materials that are not diamagnetic it is dominated by stronger effects.) Since their presence slightly *reduces* the magnetic field, diamagnetics evidently have a permeability slightly *smaller* than μ_0. Examples are water ($\mu_r = 0.9999912$), bismuth ($\mu_r = 0.99984$), and silver ($\mu_r = 0.999975$).

In other materials, atoms and molecules have a magnetic moment, but with no external magnetic field these moments are distributed randomly, and no macroscopic magnetic field results. In one class of such materials, known as *paramagnetics*, the atoms have their magnetic moments, but these moments are oriented statistically. When a field is applied, the Ampere currents of atoms align themselves with the field to some extent. This alignment is opposed by the thermal motion of the atoms, so it increases as the temperature decreases and as the applied magnetic field becomes stronger. The result of the alignment of the Ampere currents is a very small magnetic field added to the external field.

© 2006 by Taylor & Francis Group, LLC



(a) (b)

Figure 3.6 Schematic of an unmagnetized (a) paramagnetic and (b) ferromagnetic materials. The arrows show qualitatively atomic (or molecular) magnetic moments.

For paramagnetic materials, therefore, μ is slightly greater than μ_0, and μ_r is slightly greater than one. Examples are air ($\mu_r = 1.00000036$) and aluminum ($\mu_r = 1.000021$).

The most important magnetic materials in electrical engineering are known as *ferromagnetics*. They are, in fact, paramagnetic materials, but with very strong interactions between atoms (or molecules). As a result of these interactions, groups of atoms (10^{12} to 10^{15} atoms in a group) form inside the material, and in these groups the magnetic moments of all the molecules are oriented in the same direction. These groups of molecules are called *Weiss domains*. Each domain is, in fact, a small saturated magnet. A sketch of atomic (or molecular) magnetic moments in paramagnetic and ferromagnetic materials is given in Fig. 3.6.

The size of a domain varies from material to material. In iron, for example, under normal conditions, the linear dimensions of the domains are 10μm. In some cases they can get as large as a few millimeters or even a few centimeters across. If a piece of a highly polished ferromagnetic material is covered with fine ferromagnetic powder, it is possible to see the outlines of the domains under a microscope. The boundary between two domains is not abrupt, and it is called a *Bloch wall*. This is a region 10^{-8}–10^{-6}μm in width (500 to 5000 interatomic distances), in which the orientation of the atomic (or molecular) magnetic moments changes gradually.

Above a certain temperature, the *Curie temperature*, the thermal vibrations completely prevent the parallel alignment of the atomic (or molecular) magnetic moments, and ferromagnetic materials become paramagnetic. For example, the Curie temperature of iron is 770°C (for comparison, the melting temperature of iron is 1530°C).

In materials referred to as *antiferromagnetics*, the magnetic moments of adjacent molecules are antiparallel, so that the net magnetic moment is zero. (Examples are FeO, $CuCl_2$ and FeF_2, which are not widely used.) *Ferrites* are a class of antiferromagnetics very widely used at radio frequencies. They also have antiparallel moments, but, because of their asymmetrical structure, the net magnetic moment is not zero, and the Weiss domains exist. Ferrites are weaker magnets than ferromagnetics, but they have high electrical resistivities, which makes them important for high-frequency applications. Figure 3.7 shows a schematic comparison of the Weiss domains for ferromagnetic, antiferromagnetic and ferrite materials.

Ferromagnetic materials are nonlinear, i.e., $\mathbf{B} \neq \mu\mathbf{H}$. How does a ferromagnetic material behave when placed in an external magnetic field? As the external magnetic field is increased from zero, the domains that are approximately aligned with the field increase in size. Up to a certain (not large) field magnitude, this process is reversible—if the field is turned off, the domains go back to their initial states. Above a certain field strength, the domains start rotating under the influence of magnetic forces, and this process is irreversible. The domains will keep rotating until they are all aligned with the local

© 2006 by Taylor & Francis Group, LLC

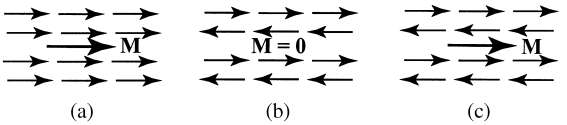


Figure 3.7 Schematic of Weiss domains for (a) ferromagnetic, (b) antiferromagnetic, and (c) ferrite materials. The arrows represent atomic (or molecular) magnetic moments.

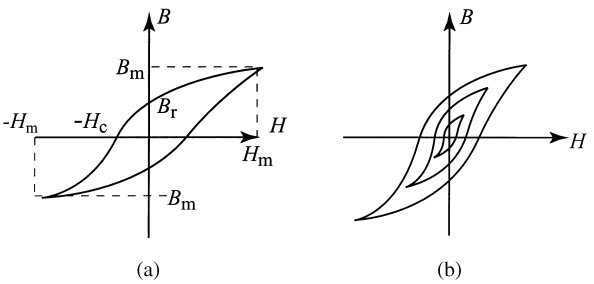


Figure 3.8 (a) Typical hysteresis loop for a ferromagnetic material. (b) The hysteresis loops for external fields of different magnitudes have different shapes. The curved line connecting the tips of these loops is known as the *normal* magnetization curve.

magnetic flux density vector. At this point, the ferromagnetic is *saturated*, and applying a stronger magnetic field does not increase the magnetization vector.

When the domains rotate, there is friction between them, and this gives rise to some essential properties of ferromagnetics. If the field is turned off, the domains cannot rotate back to their original positions, since they cannot overcome this friction. This means that some *permanent* magnetization is retained in the ferromagnetic material. The second consequence of friction between domains is loss to thermal energy (heat), and the third consequence is *hysteresis*, which is a word for a specific nonlinear behavior of the material. This is described by curves $B(H)$, usually measured on toroidal samples of the material. These curves are closed curves around the origin, and they are called *hysteresis loops*, Fig. 3.8a. The hysteresis loops for external fields of different magnitudes have different shapes, Fig. 3.8b.

In electrical engineering applications, the external magnetic field is in many cases approximately sinusoidally varying in time. It needs to pass through several periods until the $B(H)$ curve stabilizes. The shape of the hysteresis loop depends on the frequency of the field, as well as its strength. For small field strengths, it looks like an ellipse. It turns out that the ellipse approximation of the hysteresis loop is equivalent to a *complex permeability*. For sinusoidal time variation of the field, in complex notation we can write $\underline{\mathbf{B}} = \underline{\mu}\mathbf{H} = (\mu' - j\mu'')\underline{\mathbf{H}}$, where underlined symbols stand for complex quantities. (This is analogous to writing that a complex voltage equals the product of complex impedance and complex current.) This approximation does not take saturation into account. It can be shown that the imaginary part, μ'', of the complex permeability describes ferromagnetic material hysteresis losses that are proportional to frequency (see chapter on electromagnetic induction). In ferrites, which are sometimes referred to as *ceramic ferromagnetic materials*, the dielectric losses, proportional to f^2, exist in addition (and may even be dominant).

© 2006 by Taylor & Francis Group, LLC

Table 3.1 Magnetic Properties of Some Commonly Used Materials

Material	Relative permeability, μ_r	Comment
Silver	0.9999976	Diamagnetic
Copper	0.99999	Diamagnetic
Gold	0.99996	Diamagnetic
Water	0.9999901	Diamagnetic
Aluminum	1.000021	Paramagnetic
Moly permalloy	100 (few)	Ferromagnetic with air
Ferrite	1000	For example, $NiO \cdot Fe_2O_3$, insulator
Nickel	600	Ferromagnetic
Steel	2000	Ferromagnetic
Iron (0.2 impurity)	5000	Ferromagnetic
Purified iron (0.05 impurity)	2×10^5	Ferromagnetic
Supermalloy	As high as 10^6	Ferromagnetic

The ratio B/H (corresponding to the permeability of linear magnetic materials) for ferromagnetic materials is not a constant. It is possible to define several *permeabilities*, e.g., the one corresponding to the initial, reversible segment of the magnetization curve. This permeability is known as the *initial* permeability. The range is very large, from about $500\,\mu_0$ for iron to several hundreds of thousands μ_0 for some alloys.

The ratio B/H along the normal magnetization curve (Fig. 3.8b) is known as the *normal* permeability. If we magnetize a material with a dc field, and then add to this field a small sinusoidal field, a resulting small hysteresis loop will have a certain ratio $\Delta B/\Delta H$. This ratio is known as the *differential* permeability. Table 3.1 shows some values of permeability for commonly used materials.

3.2.13. Magnetic Circuits

Perhaps the most frequent and important practical applications of ferromagnetic materials involve cores for transformers, motors, generators, relays, etc. The cores have different shapes, they may have air gaps, and they are magnetized by a current flowing through a coil wound around a part of the core. These problems are hard to solve strictly, but the approximate analysis is accurate enough and easy, because it resembles dc circuit analysis.

We will restrict our attention to thin linear magnetic circuits, i.e., to circuits with thickness much smaller than their length, as in Fig. 3.9, characterized approximately by a convenient permeability (e.g., initial permeability), assumed to be independent of the magnetic field intensity. The magnetic flux in the circuit is determined from the equations.

Ampere's law applied to a contour that follows the center of the magnetic core in Fig. 3.9 can be written as

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = H_1 L_1 + H_2 L_2 = NI \tag{3.28}$$

© 2006 by Taylor & Francis Group, LLC

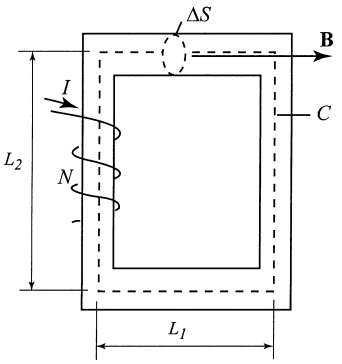


Figure 3.9 A thin magnetic circuit. L_1 and L_2 are lengths of the core sides along contour C through the center of the magnetic core of cross-section ΔS.

where

$$H_i = \frac{B_i}{\mu_i} = \frac{1}{\mu_i}\frac{\Phi_i}{S_i} \tag{3.29}$$

is the magnetic field intensity in each section of the core, assuming a linear magnetic material or a small-signal (dynamic) permeability. An additional equation is obtained for the magnetic fluxes Φ_i at the "nodes" of the magnetic circuit, recalling that

$$\oint_{S_0} \mathbf{B} \cdot d\mathbf{S} = \sum_i \Phi_i = 0 \tag{3.30}$$

for any closed surface S_0. Equations (3.28)–(3.30) can be combined to have the same form as the analogous Kirchoff's laws for electrical circuits:

$$\sum_i \Phi_i = 0 \qquad \text{for any node}$$

which is analogous to

$$\sum_i I_i = 0 \tag{3.31}$$

$$\sum_i R_{mi}\Phi_i - \sum_i N_i I_i = 0 \qquad \text{for any closed loop}$$

analogous to

$$\sum_i R_i I_i - \sum_i V_i = 0 \tag{3.32}$$

$$R_{mi} = \frac{1}{\mu_i}\frac{L_i}{S_i} \qquad \text{for any branch}$$

© 2006 by Taylor & Francis Group, LLC

analogous to

$$R_i = \frac{1}{\sigma_i}\frac{L_i}{S_i} \tag{3.33}$$

where R_m is referred to as *magnetic resistance*, and σ is the electrical conductance. The last equation is Ohm's law for uniform linear resistors.

If the magnetic circuit contains a short air gap, L_0 long, the magnetic resistance of the air gap is calculated as in Eq. (3.33), with $\mu_i = \mu_0$.

3.3. APPLICATIONS OF MAGNETOSTATICS

The sections that follow describe briefly some common applications of magnetostatic fields and forces, with the following outline:

1. Forces on charged particles (cathode ray tubes, Hall effect devices)
2. Magnetic fields and forces of currents in wires (straight wire segment, Helmholtz coils)
3. Magnetic fields in structures with some degree of symmetry (toroidal coil, solenoid, coaxial cable, two-wire line, strip-line cable)
4. Properties of magnetic materials (magnetic shielding, magnetic circuits)
5. System-level applications (magnetic storage, Magnetic Resonance Imaging—MRI).

3.3.1. Basic Properties of Magnetic Force on a Charged Particle (the Lorentz Force)

By inspecting the Lorentz force in Eq. (3.2), we come to the following conclusion: The speed of a charged particle (magnitude of its velocity) can be changed by the electric force $Q\mathbf{E}$. It *cannot* be changed by the magnetic force $Q\mathbf{v} \times \mathbf{B}$, because magnetic force is always normal to the direction of velocity. Therefore, charged particles can be accelerated only by electric forces.

The ratio of the maximal magnetic and maximal electric force on a charged particle moving with a velocity $\mathbf{v}$ equals vB/E. In a relatively large domain in vacuum, it is practically impossible to produce a magnetic flux density of magnitude exceeding 1 T, but charged particles, e.g., electrons, can easily be accelerated to velocities on the order of 1000 km/s. To match the magnetic force on such a particle, the electric field strength must be on the order of 10^6V/m, which is possible, but not easy or safe to achieve. Therefore, for example, if we need to substantially deflect an electron beam in a small space, we use magnetic forces, as in television or computer-monitor cathode-ray tubes.

The horizontal component of the earth's magnetic field is oriented along the north–south direction, and the vertical component is oriented downwards on the northern hemisphere and upwards on the southern hemisphere. Therefore, cathode-ray tubes that use magnetic field deflection have to be tuned to take this external field into account. It is likely that your computer monitor (if it is a cathode-ray tube) will not work exactly the same way if you turn it sideways (it might slightly change colors or shift the beam by a couple of millimeters) or if you use it on the other side of the globe.

© 2006 by Taylor & Francis Group, LLC

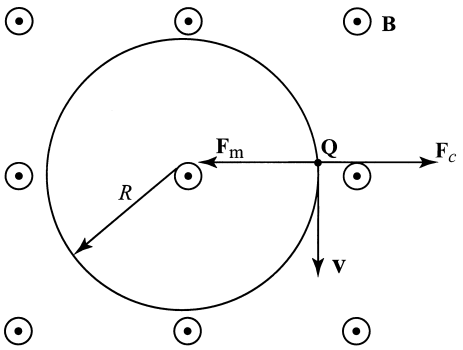


Figure 3.10 Charged particle in a uniform magnetic field.

Charged Particle Moving in a Uniform Magnetic Field

Consider a charged particle $Q > 0$ moving in a magnetic field of flux density **B** with a velocity **v** normal to the lines of vector **B**, Fig. 3.10. Since the magnetic force on the charge is always perpendicular to its velocity, it can only change the direction of the charged particle motion. To find the trajectory of the particle, note that the magnetic force on the particle is directed as indicated, tending to curve the particle trajectory. Since **v** is normal to **B**, the force magnitude is simply QvB. It is opposed by the centrifugal force, mv^2/R, where R is the radius of curvature of the trajectory. Therefore,

$$QvB = \frac{mv^2}{R} \tag{3.34}$$

so that the radius of curvature is constant, $R = mv/QB$. Thus, the particle moves in a circle. It makes a full circle in

$$t = T = \frac{2\pi R}{v} = \frac{2\pi m}{QB} \tag{3.35}$$

seconds, which means that the frequency of rotation of the particle is equal to $f = 1/T = QB/2\pi m$. Note that f does not depend on v. Consequently, all particles that have the same charge and mass make the same number of revolutions per second. This frequency is called the *cyclotron* frequency. Cyclotrons are devices that were used in the past in scientific research for accelerating charged particles. A simplified sketch of a cyclotron is shown in Fig. 3.11, where the main part of the device is a flat metal cylinder, cut along its middle. The two halves of the cylinder are connected to the terminals of an oscillator (source of very fast changing voltage). The whole system is in a uniform magnetic field normal to the bases of the cylinder, and inside the cylinder is highly rarefied air.

A charged particle from source O finds itself in an electric field that exists between the halves of the cylinder, and it accelerates toward the other half of the cylinder. While outside of the space between the two cylinder halves, the charge finds itself only in a magnetic field, and it circles around with a radius of curvature $R = mv/QB$. The time it takes to go around a semicircle does not depend on its velocity. That means that it will

© 2006 by Taylor & Francis Group, LLC

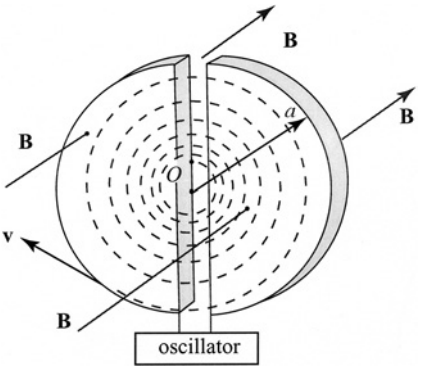


Figure 3.11 Sketch of a cyclotron used in scientific research to accelerate charged particles by means of electric and magnetic fields.

always take the charge the same amount of time to reach the gap between the two cylinders. If the electric field variation in this region is adjusted in such a way that the charge is always accelerated, the charge will circle around in larger and larger circles, with increasingly larger velocity, until it finally shoots out of the cyclotron. The velocity of the charge when it gets out of the cyclotron is $v = QBa/m$. This equation is valid only for velocities not close to the speed of light. For velocities close to the speed of light, the mass is not constant (it is increased due to relativistic effects). As a numerical example, for $B = 1\,\text{T}$, $Q = e$, $a = 0.5\,\text{m}$, $m = 1.672 \times 10^{-27}\,\text{kg}$ (a proton), the particles will be accelerated to velocities $v = 47.9 \times 10^6\,\text{m/s}$. Cyclotrons are not used any more for particle physics research, but they were instrumental in the initial phases of this science. At the University of Chicago, for example, a cyclotron was used for research that led to the development of the atomic bomb.

The Hall Effect

In 1879, Edwin Hall discovered an effect that can be used for measuring the magnetic field and for determining the sign of free charges in conductors. Let a conducting ribbon have a width d and is in a uniform magnetic field of flux density **B** perpendicular to the ribbon, Fig. 3.12. A current of density **J** flows through the ribbon. The free charges can in principle be positive (Fig. 3.12a) or negative (Fig. 3.12b). The charges that form the current are moving in a magnetic field, and therefore there is a magnetic force $\mathbf{F} = Q\mathbf{v} \times \mathbf{B}$ acting on them. Due to this force, positive charges accumulate on one side of the ribbon, and negative ones on the other side. These accumulated charges produce an electric field E_H. This electric field, in turn, acts on the free charges with a force that is in the opposite direction to the magnetic force. The charges will stop accumulating when the electric force is equal in magnitude to the magnetic force acting on each of the charges. Therefore, in steady state

$$QvB = QE_H \qquad \text{or} \qquad E_H = vB \tag{3.36}$$

Between the left and right edge of the ribbon, one can measure a voltage equal to

$$|V_{12}| = E_H d = vBd \tag{3.37}$$

© 2006 by Taylor & Francis Group, LLC

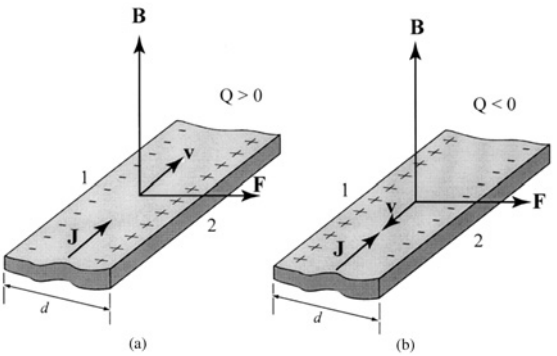


Figure 3.12 The Hall effect in case of (a) positive free-charge carriers, and (b) negative free-charge carriers.

In the case shown in Fig. 3.12a, this voltage is negative, and in Fig. 3.12b it is positive. Thus, the sign of the voltage corresponds to the sign of free charge carriers and can be determined by a voltmeter.

Since $J = NQv$, where N is the number of free charges per unit volume, the Hall voltage becomes

$$|V_{12}| = \frac{Jd}{NQ}B \tag{3.38}$$

Thus, if the coefficient Jd/NQ is determined for a ribbon such as the one sketched in Fig. 3.12, by measuring V_{12}, the magnetic flux density B can be determined. Usually, Jd/NQ is determined experimentally. This ribbon has four terminals: two for the connection to a source producing current in the ribbon, and two for the measurement of voltage across it. Such a ribbon is called a *Hall element*.

For single valence metals, e.g., copper, if we assume that there is one free electron per atom, the charge concentration is given by

$$N = \frac{N_A \rho_m}{M} \tag{3.39}$$

where N_A is Avogadro's number (6.02×10^{23}atoms/mol), ρ_m is the mass density of the metal, and M is the atomic mass.

As a result of the above properties, Hall elements are key components in devices used for a wide range of measurements:

The Hall effect is most pronounced in semiconductors. Hall-effect devices are commonly used to determine the type and concentration of free carriers of semiconductor samples, as can be deduced from Eqs. (3.38) and (3.39).

Gaussmeters (often called *teslameters*) use a Hall element to measure magnetic flux density, by generating output voltage proportional to the magnetic field. Special attention is given to the design of the accompanying Hall-effect probes. The accuracy and calibration of Hall-effect Gaussmeters is verified by standardized reference magnets.

© 2006 by Taylor & Francis Group, LLC

In integrated circuits technology, the Hall effect is used for sensors and switches. In sensors, the magnetic flux density through the Hall element determines the output voltage; in switches, it determines the switching state. Hall-effect sensor operation is robust with respect to environmental conditions.

Linear Hall sensors, which generate voltage proportional to the magnetic flux perpendicular to the Hall plate, are characterized by output quiescent voltage (the output voltage in absence of the magnetic field) and sensitivity. Their industrial applications include measurement of angle, current, position and distance, and pressure, force, and torque sensors. In automotive industry, they are used for active suspension control, headlight range adjustment, liquid level sensors, power steering, and so on. With very low energy consumption (a fraction of a mW), linear Hall sensors are more efficient and cost effective than most inductive and optoelectronic sensors.

A Hall switch contains an integrated comparator with predefined switching levels and an open-drain transistor at its digital output, which can be adapted to different logic systems. The output characteristic of a Hall switch resembles a hysteresis-like (B, V_{out}) curve. The magnetic flux density B of the hysteresis ranges from B_{off} to B_{on}; if $B > B_{on}$, the output transistor is switched on, and if $B < B_{off}$, the transistor is switched off. These switches are also available in a differential form, where the output transistor is switched according to the difference of the magnetic flux between two Hall-element plates separated typically by several millimeters. Finally, in the case of two-wire Hall switches, the output signal of the switch is a current of an internal source, which is switched on or off by the magnetic field applied to the Hall plate. In all Hall switches, simplified switching ensures a clean, fast, and bounceless switch avoiding the problems present in mechanical contact switches. Hall-effect switches are more cost effective than most electromechanical switches. Among other applications, they are widely used for commutation of brushless DC motors, wheel speed sensors, measurement of rotations per minute, pressure switches, position-dependent switches, etc. The automotive industry uses Hall switches, e.g., in ignition and wiper systems, door locks, window raising controls, and retraction-roof controls and for break light switches. In the computer industry, this type of switch is used in keyboards.

3.3.2. Magnetic Fields of Currents in Wires

Biot–Savart's law in Eq. (3.3) can be used to calculate vector **B** produced by currents in wire loops of arbitrary shapes (i.e., a variety of electrical circuits). Such loops are often made of (or can be approximated by) a sequence of interconnected straight wire segments. Evaluation of **B** in such cases can greatly be simplified if we determine vector **B** produced by the current in a single straight wire segment. With reference to Fig. 3.13, using Biot–Savart's law, the following expression is obtained

$$B = \frac{\mu_0 I}{4\pi a}(\sin\theta_2 - \sin\theta_1) \tag{3.40}$$

Helmholtz Coils

To obtain in a simple manner highly uniform magnetic field in a relatively large domain of space in air, *Helmholtz coils* can be used. They consist of two thin, parallel, coaxial circular loops of radius a that are a distance a apart, Fig. 3.14a. Each loop carries a current I,

© 2006 by Taylor & Francis Group, LLC

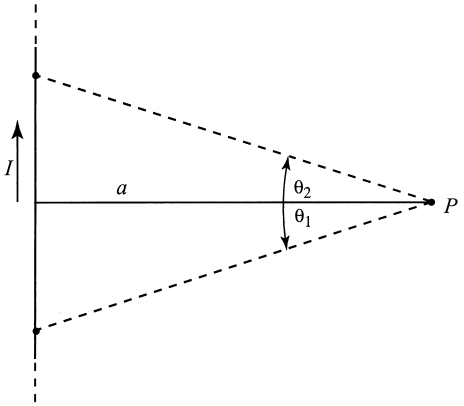
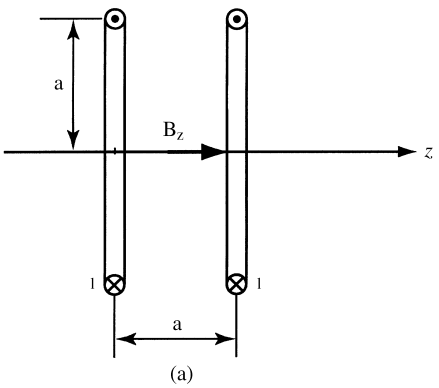


Figure 3.13 Calculating the magnetic field at point *P* due to a straight wire segment with current *I*.



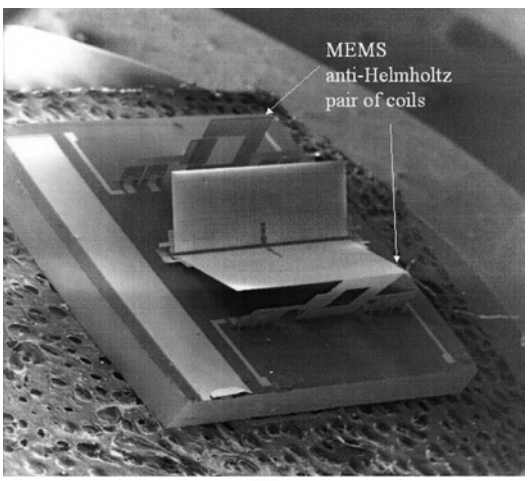


Figure 3.14 (a) Sketch of a Helmholtz pair of coils. The magnetic field in the center is highly uniform. (b) Photograph of a micro-electromachined anti-Helmholtz coils (courtesy Profs. Victor Bright and Dana Anderson, University of Colorado at Boulder). The inductors are released to spring up into position.

© 2006 by Taylor & Francis Group, LLC

and the two currents are in the same direction. Starting from the Biot–Savart law, we find that the magnetic field at a distance z from the center of one loop, on their common axis, is axial, of magnitude

$$B_z(z) = \frac{\mu_0 I a^2}{2} \frac{1}{\left[a^2 + (a-z)^2\right]^{3/2}} \tag{3.41}$$

It can be proven that, at $z = a/2$, the first, second, and even the third derivatives of $B_z(z)$ are zero, which means that the magnetic flux density around that point is highly uniform.

If the currents in Helmholtz coils are flowing in opposite directions, the magnetic field has a null in the center, accompanied by a very large gradient. An interesting application of this anti-Helmholtz pair of coils is in the emerging field of atomic optics, where large gradients of the magnetic field are used to guide atoms and even Bose-Einstein condensates. A photograph of a micro-electromachined (MEM) anti-Helmholtz pair is shown in Fig. 3.14b.

Magnetic Force Between Two Long Parallel Wires: a Definition of the Ampere

Two parallel current-carrying wires can either attract or repel each other, depending on the direction of the currents in them. If the wires are in vacuum (air) and are very long (theoretically infinitely long), if currents in them are of equal magnitude I and the distance between them is d, the force per unit length of the wires is

$$F_{\text{per unit length}} = \mu_0 \frac{I^2}{2\pi d} \tag{3.42}$$

To get a feeling for the magnetic forces between current-carrying conductors, from this equation we find that, for $d = 1$ m and $I = 1$ A, the force on each of the wires is 2×10^{-7} N/m. This used to be one of the definitions of the unit for electrical current, the *ampere*.

Magnetic Force on the Short Circuit of a Two-Wire Line

As another example, the magnetic force on the segment A-A' of the two-wire-line short circuit shown in Fig. 3.15a is given by

$$F = \frac{\mu_0 I^2}{2\pi} \ln \frac{d-a}{a} \tag{3.43}$$

If a large current surge occurs in the line, the force shock on the short circuit can be quite large. For example, if there is a sudden increase of current intensity to $I = 5000$ A, $a = 0.5$ cm, and $d = 50$ cm, the force shock is about 23 N, which may be sufficient to open the short circuit if it is not firmly connected.

Magnetic Force in a Loudspeaker

Shown in Fig. 3.15b is a sketch of a permanent magnet used in loudspeakers. The lines of the magnetic flux density vector are radial, and at the position of the coil the magnitude

© 2006 by Taylor & Francis Group, LLC

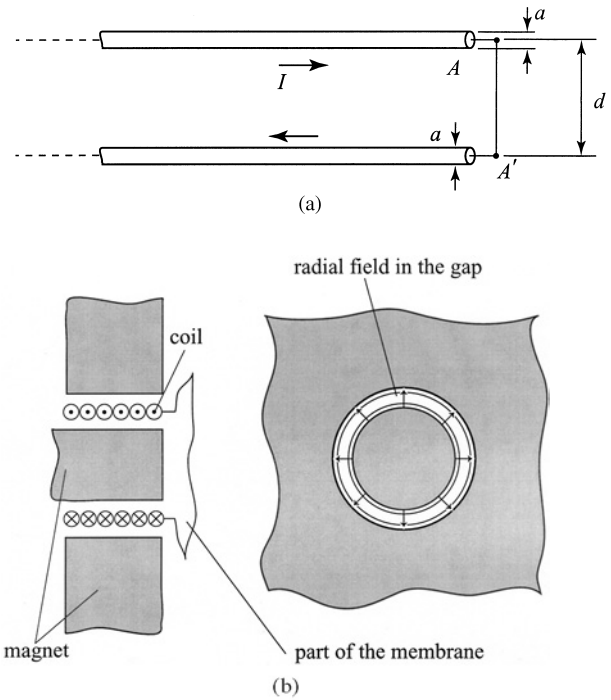


Figure 3.15 (a) A Short-circuited two-wire line. (b) Permanent magnet and coil used in a loudspeaker.

is $B = 1\,\text{T}$. Let the coil have N turns, its radius be a, and the current in the coil be I. The magnetic force on the coil (which is glued to the loudspeaker membrane) is $F = 2\pi aNIB$. If, in particular, $I = 0.15\,\text{A}$, $N = 10$, and $a = 0.5\,\text{cm}$, we find that $F = 0.047\,\text{N}$.

3.3.3. Applications of Ampere's Law

Ampere's law can be used to determine the magnetic field produced by currents in structures with a high level of symmetry. Common and practical examples are discussed below.

Magnetic Field of a Straight Wire

Consider a *straight*, very long (theoretically infinite) wire of circular cross section of radius a, Fig. 3.16a. (A wire may be considered infinitely long if it is much longer than the shortest distance from it to the observation point.) There is a current of intensity I in the wire distributed uniformly over its cross section. Note that, due to symmetry, both outside and inside the wire the lines of vectors **B** and **H** are circles centered along the wire axis and in planes normal to it. Therefore, the only unknown is the magnitude of these vectors as a function of the distance r from the wire axis. Using Ampere's law we find that

$$B(r) = \frac{\mu_0 I}{2\pi r} \qquad \text{for } r \geq a \tag{3.44}$$

© 2006 by Taylor & Francis Group, LLC

As long as the point is outside the wire, the radius of the wire a is irrelevant. This expression for B outside a round wire is valid for a wire of any radius, including an infinitely thin one. Inside the wire, the magnetic flux density is given by

$$B(r) = \frac{\mu_0 I r}{2\pi a^2} \qquad \text{for } r \le a \tag{3.45}$$

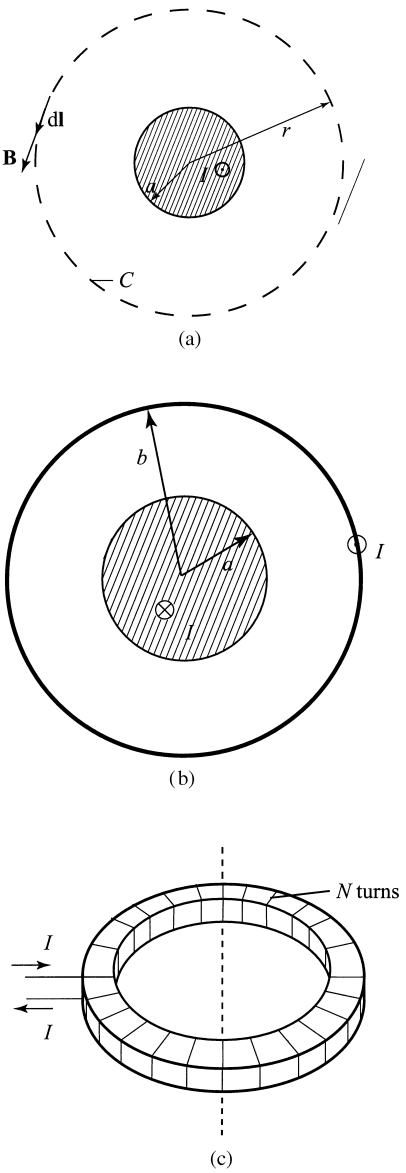


Figure 3.16 (a) Cross section of straight wire of circular cross section with a current of intensity I. (b) The cross section of a coaxial cable with very thin outer conductor. (c) A toroidal coil with N windings. (d) Longitudinal and transverse cross sections of a solenoid. (e) A current sheet. (f) Two parallel current sheets.

© 2006 by Taylor & Francis Group, LLC

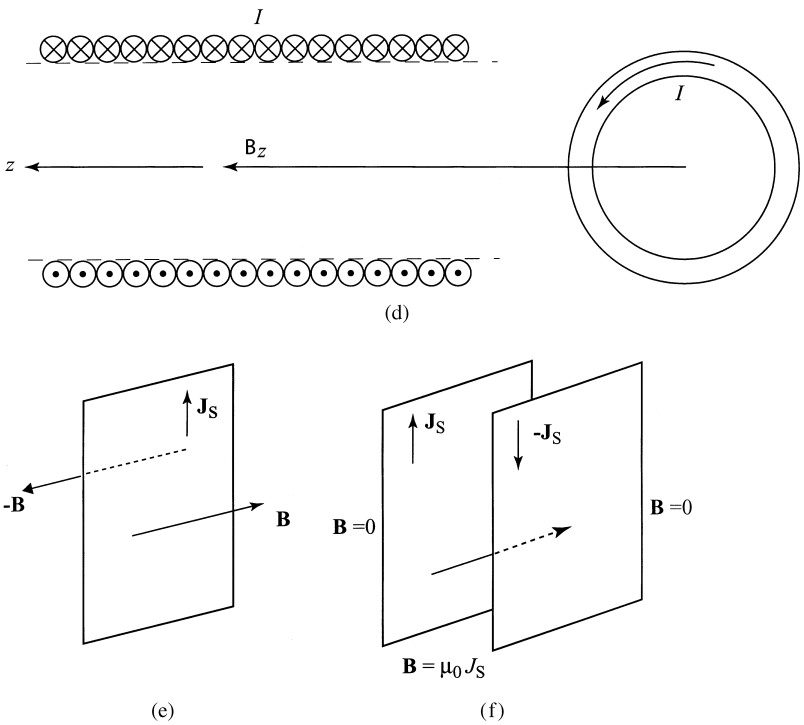


Figure 3.16 Continued.

Magnetic Field in a Coaxial Cable

Using the same procedure we find the magnetic flux density due to currents I and $-I$ in conductors of a coaxial cable, Fig. 3.16b. If the outer conductor is assumed to be very thin (as it usually is), outside the cable the magnetic field does not exist, and inside the cable, the magnetic flux density is radial and equal to

$$B(r) = \begin{cases} \dfrac{\mu_0 I r}{2\pi a^2} & \text{for } r \le a \\ \dfrac{\mu_0 I}{2\pi r} & \text{for } r \ge a \end{cases} \tag{3.46}$$

Magnetic Field of a Toroidal Coil

Another commonly used case in inductors and transformers is that of a toroidal coil, Fig. 3.16c. The cross section of the toroid is arbitrary. Assume that the coil is made of N uniformly and densely wound turns with current of intensity I. From the Biot–Savart law, we know that the lines of vector **B** are circles centered on the toroid axis. Also, the magnitude of **B** depends only on the distance r from the axis. Applying Ampere's law yields the following expressions for the magnitude, $B(r)$, of the magnetic flux density vector:

$$B = \begin{cases} 0 & \text{outside the toroid} \\ \dfrac{\mu_0 N I}{2\pi r} & \text{inside the toroid} \end{cases} \tag{3.47}$$

© 2006 by Taylor & Francis Group, LLC

As a numerical example, for $N = 1000$, $I = 2\,\text{A}$, and a mean toroid radius of $r = 10\,\text{cm}$, we obtain $B = 4\,\text{mT}$. This value can be larger if, for example, several layers of wire are wound one on top of each other, so that N is larger. Alternatively, the torus core can be made of a ferromagnetic material, resulting in much larger magnitude of the magnetic flux density inside the core.

Magnetic Field of a Long Solenoid

Assume that the radius r of the toroid becomes very large. Then, at any point inside the toroid, the toroid looks locally as if it were a cylindrical coil, known as a *solenoid* (from a Greek word which, roughly, means "tubelike"), Fig. 3.16d. We conclude that outside an "infinitely long" solenoid the flux density vector is zero. Inside, it is given by Eq. (3.47) with r very large, or since $N' = N/2\pi r$ is the number of turns per unit length of the toroid, i.e., of the solenoid,

$$B = \mu_0 N' I \qquad \text{inside the solenoid (coil)} \tag{3.48}$$

The field inside a very long solenoid is *uniform*, and the expression is valid for *any* cross section of the solenoid. As a numerical example, $N' = 2000$ windings/m and $I = 2\,\text{A}$ result in $B \cong 5\,\text{mT}$.

Magnetic Field of a Planar Current Sheet and Two Parallel Sheets

Consider a large conducting sheet with constant surface current density $\mathbf{J}_s$ at all points, Fig. 3.16e. From the Biot–Savart law, vector $\mathbf{B}$ is parallel to the sheet and perpendicular to vector $\mathbf{J}_s$, and $\mathbf{B}$ is directed in opposite directions on the two sides of the sheet, as indicated in the figure. Applying Ampere's law gives

$$B = \mu_0 \frac{J_S}{2} \qquad \text{for a current sheet} \tag{3.49}$$

For two parallel current sheets with opposite surface currents of the same magnitude (Fig. 3.16f), from the last equation and using superposition we find that the magnetic field outside the sheets is zero, and between two parallel current sheets

$$B = \mu_0 J_S \tag{3.50}$$

Magnetic Field of a Stripline

Equation (3.50) is approximately true if the sheets are not of infinite width, and are close to each other. Such a system is called a *strip line*. Assume that the strip line is filled with a ferrite of relative permeability μ_r. Since $a \gg g$, where a is the finite strip width and g is the distance (gap) between two infinitely long strips, the magnetic field outside the strips can be neglected, and the resulting magnetic flux density inside the strip line is $B = \mu_r \mu_0 I/a$. The magnitude of the magnetization vector in the ferrite is $M = (\mu_r - 1)I/a$. The density of equivalent surface magnetization currents is thus $J_{ms} = (\mu_r - 1)I/a$. These currents have the same direction as the conduction currents in the strips, but are many times greater than the surface current over the strips.

© 2006 by Taylor & Francis Group, LLC

3.3.4. Magnetic Shielding; Magnetic Materials for EMC Testing

Imagine two cavities (air gaps) inside an uniformly magnetized material of relative permeability μ_r. One is a needlelike cavity in the direction of the vector **B**. The other is a thin-disk cavity, normal to that vector. According to boundary conditions, the ratio of magnitudes of the magnetic flux density vectors in the two cavities and that in the surrounding material is equal to $1/\mu_r$ and 1, respectively. We can therefore conclude that the theoretical possibility of reducing the external time-invariant magnetic field by means of "magnetic shielding" is by a factor of $1/\mu_r$. Note, however, that the shielding effect of *conductive* ferromagnetic materials is greatly increased for time-varying fields, due to the skin effect (see chapter on electromagnetic induction). Note that, for EMC/EMI (electromagnetic compatibility and electromagnetic interference) testing, ferrite anechoic chambers are used. These rely on magnetic losses inside ferrite materials and will briefly be discussed in Chapter 4.

3.3.5. Measurements of Basic Properties of Magnetic Materials

The curve $B(H)$ that describes the nonlinear material is usually obtained by measurement. The way this is done is sketched in Fig. 3.17. A thin toroidal core of mean radius R, made of the material we want to measure, has N tightly wound turns of wire, and a cross-sectional area S. If there is a current I through the winding, the magnetic field intensity inside the core is given by

$$H = \frac{NI}{2\pi R} \tag{3.51}$$

From this formula, the magnetic field magnitude for any given current can be calculated. Around the toroidal core there is a second winding, connected to a ballistic galvanometer (an instrument that measures the charge that passes through a circuit). It can be shown that the charge that flows through the circuit is proportional to the change of the magnetic flux, $\Delta Q \propto \Delta\Phi = S\,\Delta B$, and therefore to the change of the B field as well. By changing the current I through the first winding, the curve $B(H)$ can be measured point by point. If the field H is changing slowly during this process, the measured curves are called *static magnetization curves*.

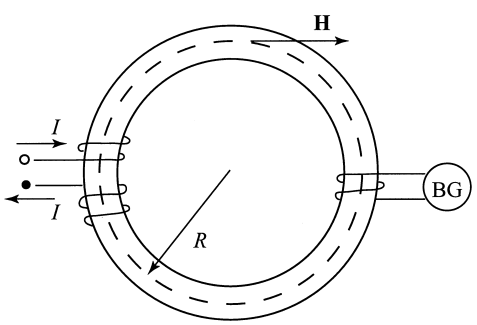


Figure 3.17 Sketch of setup for measurement of magnetization curves.

© 2006 by Taylor & Francis Group, LLC

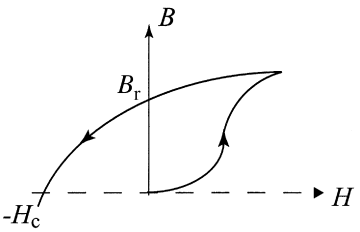


Figure 3.18 Magnetization and demagnetization curves.

Permanent Magnets

If a ferromagnetic body is magnetized (desirably to saturation) and the external magnetic field is switched off, the body remains magnetized, i.e., becomes a permanent magnet. If the body is a thin toroid and the magnetic field is produced by a uniformly wound coil, the magnetization curve is as in Fig. 3.18. When the current after saturation of the toroid is switched off, the operating point moves to the point labeled B_r, known as the *remanent* flux density. If the current is reversed (changes sign), the point moves along the curve to the left of the B axis, sometimes referred to as the demagnetization curve. The magnetic field intensity H_c corresponding to zero B is known as the *coercive* magnetic field. If we cut a piece out of the magnetized toroid with remanent flux density inside it, the operating point will move along the demagnetization curve. A magnetic field will exist in the air gap, and a permanent toroidal magnet is obtained.

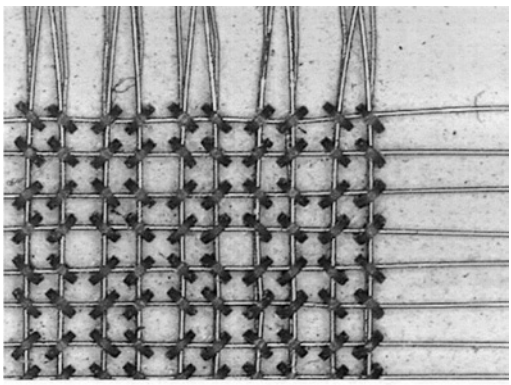
Permanent magnets are used in a large variety of applications, one of the most common being data storage on magnetic disks and tapes, in the form of small magnetized patches on thin magnetic films deposited on plastic substrate.

3.3.6. Magnetic Storage

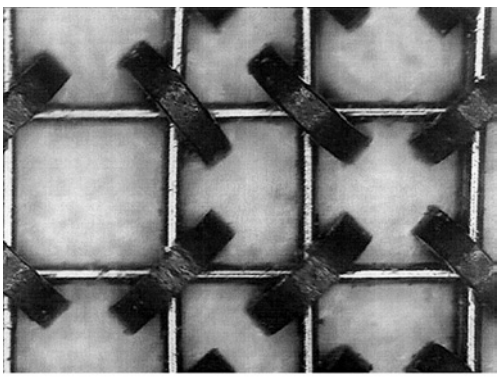
Magnetic materials have been used for storing data since the very first computers. In the 1970s, magnetic core memories were used and an example is shown in Fig. 3.19. The principle of operation of magnetic core memories is an excellent illustration of both magnetostatics and electromagnetic induction. Furthermore, it appears that these components might see a revival for space applications due to the radiation hardness of magnetic materials. In Fig. 3.19a, one bit of the memory is a small ferromagnetic torus, with two wires passing through it. One of the wires is connected to a circuit used for both writing and reading, and the second wire is used only for reading. To write a "1" or a "0", a positive or negative current pulse, respectively, is passed through the first wire. This results in the core magnetized to either B_r or $-B_r$ on the hysteresis curve, respectively. Elements of an entire memory are arranged in matrices, with two wires passing through each torus, Fig. 3.19b. The current passing through each row or column is half of the current needed to saturate the torus, so both the row and the column of the specific bit need to be addressed. The readout process requires electromagnetic induction and will be described in the next chapter.

A common magnetic storage device used today is the hard disk drive in every computer. Information is written to the disk by magnetizing a small piece of the disk surface. As technology is improving, the amount of information that can be stored on a

© 2006 by Taylor & Francis Group, LLC



(a)



(b)

Figure 3.19 (a) A portion of a magnetic core memory matrix and (b) a close-up showing individual memory elements, where each torus represents 1 bit. The wire radius is 35 μm and the core diameter 420 μm.

standard-size hard disk is rapidly growing [6,14]. In 2002, drives with more than 20 Gbytes were readily available in personal computers, while in 1995, a few hundred mega bytes were standard. The development is in the direction of increasing disk capacity and increasing speed (or reducing access time). These two requirements compete with each other, and the engineering solution, as is usually the case, needs to be a compromise.

We will now describe in a simple manner how data are written on the disk. The hard disk itself is coated with a thin coating of ferromagnetic material such as Fe_2O_3. The disk is organized in sectors and tracks, as shown in Fig. 3.20a.

The device that writes data to the disk (and reads data from it) is called a magnetic head. Magnetic heads are made in many different shapes, but all operate essentially according to the same principle. We here describe the operation of the writing process on the example of a simplified head which is easy to understand and is shown in Fig. 3.21. The head is a magnetic circuit with a gap. The gap is in close proximity to the tracks, so there is some leakage flux between the head and the ferromagnetic track.

In the "write" process, a current flows through the windings of the magnetic head, thus creating a fringing magnetic field in the gap. The gap is as small as 5 μm. As the head

© 2006 by Taylor & Francis Group, LLC

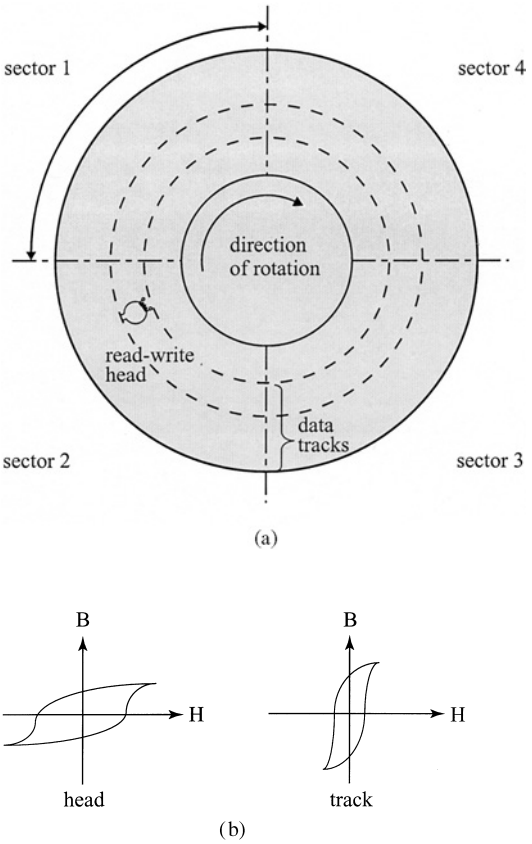


Figure 3.20 (a) Hard disk tracks. (b) Sketch of qualitative shapes of hysteresis curves required for the head and track magnetic materials.

moves along the track (usually the track rotates), the fringing field magnetizes a small part of the track, creating a south and a north pole in the direction of rotation. These small magnets are about 5 μm long and 25 μm wide. A critical design parameter is the height of the head above the track: the head should not hit the track, but it also needs to be as close as possible to maximize the leakage flux that magnetizes the track. Typically, the surface of the track is flat to within several micrometers, and the head follows the surface profile at a distance above it of about 1 μm or less. This is possible because the head aerodynamically flies above the disk surface. The current in the head windings should be strong enough to saturate the ferromagnetic track. If the track is saturated and the remanent flux density of the track material is large, the voltage signal during readout is maximized. The requirements on material characteristics for the head and tracks are completely opposite: the head needs to have a low value of the remanent magnetic flux density, since during readout any remaining B essentially represents noise. In contrast, the track material needs to stay magnetized as long as possible with as high a B as possible. A sketch of the relative desired hysterisis curves is shown in Fig. 3.20b. The principle of readout is an excellent example of electromagnetic induction and is described briefly in the next chapter.

© 2006 by Taylor & Francis Group, LLC

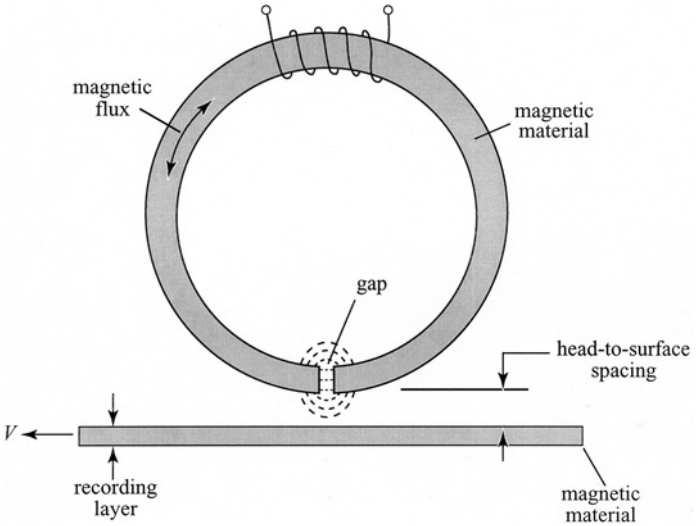


Figure 3.21 The magnetic head aerodynamically flies over the disk surface at a distance above it of only about 1μm while following the surface profile. In the figure, the surface profile is shown as ideally flat, which in practice is not the case.

3.3.7. Magnetic Circuits

Consider a thin toroidal coil of length L, area of cross section S, and with N turns. Assume that the permeability of the core is μ and that a current I is flowing through the coil. Using Eqs. (3.32) and (3.33), the following is obtained

$$\Phi = \frac{NI}{R_m} = \frac{NI}{(L/\mu S)} = \mu \frac{N}{L} IS = \mu N' IS \tag{3.52}$$

This is the same result as that obtained by determining B for the coil using Ampere's law, and $\Phi = B \cdot S$.

The analysis of arbitrarily complex thin linear magnetic circuits is very simple—it is analogous to the analysis of dc electrical circuits. However, real magnetic circuits are neither thin, nor linear. Nevertheless, thin linear magnetic circuits can be used as the basis for approximate analysis of actual magnetic circuits.

Consider a thick, U-shaped core of permeability $\mu_1 \gg \mu_0$, closed by a thick bar of permeability $\mu_2 \gg \mu_0$, as shown in Fig. 3.22. N turns with a current I are wound on the core. The exact determination of the magnetic field in such a case is almost impossible. The first thing we can conclude is that since $\mu_1, \mu_2 \gg \mu_0$, the tangential component of the magnetic flux density **B** is much larger in the core than in the air outside it. The normal components of **B** are equal, so the magnetic flux density inside the core is generally much larger than outside the core. Therefore, the magnetic flux can be approximately considered to be restricted to the core. This is never exactly true, so this is the first assumption we are making.

Further, if we assume that Eqs. (3.32) and (3.33) are reasonably accurate if lengths L_1 and L_2 are used as average lengths for the two circuit sections (with their actual cross-sectional areas), we can approximately analyze the circuit using thin-circuit theory. It is instructive to show that the error in doing so is acceptable.

© 2006 by Taylor & Francis Group, LLC

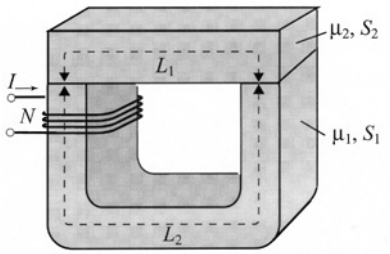


Figure 3.22 A realistic thick magnetic circuit of an inductor.

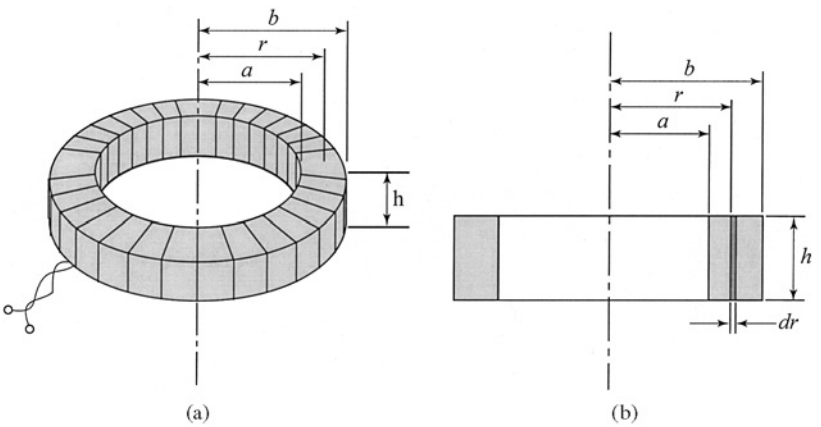


Figure 3.23 (a) A toroidal coil. (b) Cross section of the coil.

A toroidal coil and its cross section are shown in Fig. 3.23. Since the coil has N densely wound turns with a current I, from Ampere's law we find that $B = \mu NI/2\pi r$. The exact value of the magnetic flux through the toroid cross section is

$$\Phi_{\text{exact}} = \frac{\mu NIh}{2\pi}\int_a^b \frac{dr}{r} = \frac{\mu NIh}{2\pi}\ln\frac{b}{a} \tag{3.53}$$

According to Eqs. (3.32) and (3.33), adopting the average length of the toroidal core, the approximate flux is

$$\Phi_{\text{approximate}} = \frac{NI}{R_m} = \frac{NI}{(\pi(a+b))/(\mu(b-a)h)} = \frac{\mu NIh}{2\pi}\frac{2(b-a)}{b+a} \tag{3.54}$$

The relative error is

$$\frac{\Phi_{\text{approximate}} - \Phi_{\text{exact}}}{\Phi_{\text{exact}}} = \frac{2(b/a-1)}{(b/a)\ln(b/a)} - 1 \tag{3.55}$$

© 2006 by Taylor & Francis Group, LLC

which is very small even for quite thick toroids. For example, if $b/a = e = 2.718\ldots$, the error is less than 8%. Therefore, the magnetic flux in the magnetic circuit in Fig. 3.23 can be determined approximately as

$$\Phi \approx \frac{NI}{L_1/(\mu_1 S_1) + L_2/(\mu_2 S_2)} \tag{3.56}$$

If the magnetic material of a circuit cannot be approximated as linear, i.e., there is no equivalent relative permeability, the measured relationship $B(H)$ must be used.

3.3.8. Nuclear Magnetic Resonance (NMR) and Magnetic Resonance Imaging (MRI)

Superconducting loops can carry currents of enormous densities on the order of 1000 A/mm^2 and consequently can be used to make the strongest electromagnets known. Extremely strong superconducting magnets (0.5–30 T) are used in nuclear magnetic resonance (NMR) systems, best known in medical applications as magnetic resonance imaging (MRI). These devices are able to resolve three-dimensional molecule structures. Currently, NMR-based products are used in diverse fields, such as biomedical imaging, human genome research, and pharmaceutical industry [4,5,8,20].

First observed by Felix Bloch and Edward M. Purcell in 1946, the phenomenon which serves as the basis of the NMR technology can be explained as follows [1]. Nuclei of certain common atoms, such as hydrogen and carbon, have a magnetic moment of their own (referred to as *spin*). When in a strong static magnetic field, the atom spins align themselves either against or along the external magnetic field. If, in addition, a radio-frequency magnetic field is applied at exactly the magnetic-field intensity-dependent spin resonant frequency, the spin changes, producing a resonant energy state switching, which results in absorption or emission of energy. Atoms of different elements have different resonance frequencies at which the spin change occurs in the presence of a magnetic field of specified strength. This "signature frequency" allows researchers to identify the atoms and molecules present in the material under test.

Stronger magnetic fields result in increased sensitivity, permitting the analysis of smaller structures and therefore a higher resolution. The increase in the magnetic field strength results in higher concentration of the aligned spins and in higher signature resonant frequency. These two factors give rise to improved resolution by means of a higher signal-to-noise ratio. Finally, since the energy change of the spins through a single scan is very small, a clear, high signal-to-noise ratio image is achieved by superposition of many repeated NMR scans.

In most NMR systems, the strong magnetic field is produced by *superconducting* electromagnets. In some configurations, *hybrid* magnets are used, where an inner layer constructed of a resistive electromagnet is surrounded by a superconducting magnet layer. In both cases, the magnet is commonly placed in the ground, with a conveniently constructed access to the bore. Several key terms are associated with the NMR technology (Dr. Vesna Mitrović, Centre National de la Recherche Scientifique, Grenoble, France, personal communication, 2002; Dr. Mitrović is now with Brown University), and are briefly outlined next.

Spectral resolution of the NMR measurement is expressed in *parts per million* (ppm), with reference to the frequency of the radio signal used for inducing the resonance. The *bore* of the NMR magnet is the hollow part of the NMR device, which holds the

© 2006 by Taylor & Francis Group, LLC

material or body under test. The entire NMR system of magnets and coils is cooled in a pool of liquid helium. The *cold bore* structure refers to the configuration where the magnet and the coils are placed directly into the liquid helium, while the "warm bore" configuration has additional layers of vacuum and liquid nitrogen, allowing the space within the bore itself to be at the room temperature. The *shim coils* (or bobbins), the inductive coils strategically placed and current-sourced with respect to the magnet, can be found in all NMR devices and serve for tunable compensation and improvement of magnetic field homogeneity. For detailed specifications and new solutions, the reader is encouraged to read additional information available on the internet sites of the leading manufacturers and research groups: Varian Inc.; Oxford Instruments, UK; National High Magnetic Field Laboratory in Florida, U.S.A.; and Centre National de la Recherche Scientifique in Grenoble, France.

Today, NMR is an essential tool for the discovery and development of pharmaceuticals [8]. Special state-of-the-art sensitive NMR systems with high resolution used in human genome research allow structural analysis to analyze DNA samples found, for example, in protein membranes.

Since the early 1980s, NMR techniques have been used for medical visualization of soft body tissues. This application of NMR is called *magnetic resonance imaging* (MRI), and it is enabled by hydrogen nuclei present in the water and lipid content of animal tissue. Imaging magnets for animal imaging commonly have higher field strengths (3–7 T) than those used for human diagnosis (0.3–1.5 T). MRI provides high-contrast images between different tissues (brain, heart, spleen, etc.) and is sufficiently sensitive to differentiate between normal tissues and those that are damaged or diseased. *Functional MRI* (fMRI) [20] uses higher-field magnets (4 T) to help visualize the activity of the sensory, cognitive and motor system. Figure 3.24 shows an example of an MRI scan of the brain of one of



Figure 3.24 An MRI scan of the brain of one of the authors, performed using a GE instrument with a magnetic field flux density of 1.5 T (courtesy University of Colorado Health Science Center, Denver, Colorado).

© 2006 by Taylor & Francis Group, LLC

the authors. At the time of this writing, the main manufacturers or MRI imaging systems are Siemens and General Electric.

REFERENCES AND FURTHER READING

1. Bloch, F.; Hansen, W.W.; Packard, M. The nuclear induction experiment. Phys. Rev. *70*, 474–485.
2. Carter, G.W. *The Electromagnetic Field in Its Engineering Aspects*; American Elsevier: New York, 1967.
3. Cheng, D. K. *Fundamentals of Engineering Electromagnetics*; Addison-Wesley, Reading, MA, 1993, pp. 172–194.
4. Damadian, D.V. Tumor detection by nuclear magnetic resonance. Science, **March**, **1971**.
5. Freeman, R. *A Handbook of Nuclear Magnetic Resonance*; Longman Scientific & Technical: Essex, England, 1988.
6. Grochowski E. The Continuing Evolution of Magnetic Hard Disk Drives. In *Holographic Data Storage*, Coufal H.J., Psaltis D., Sincerbox G.T. Eds.; Springer Series in Optical Sciences, Springer-Verlag; Berlin, 2000; 447–462.
7. Guru, B.S.; Hiziroglu, H.R. *Electromagnetic Field Theory Fundamentals*; PWS Publishing Company: Boston, 1998, pp. 155–187.
8. Gunther, H. *NMR Spectroscopy: Basic Principles*, *Concepts*, *and Applications in Chemistry*, 2nd Ed.; John Wiley & Son: New York, 1995.
9. Harrington, R.F. *Introduction to Electromagnetic Engineering*; McGraw-Hill: New York, 1958.
10. Hoole, S.R.H.; Hoole, R.R.P. *A modern Short Course in Engineering Electromagnetics*; Oxford University Press: New York, 1996, pp. 246–260.
11. Jackson, J.D. *Classical Electrodynamics*; John Wiley & Sons: New York, 1962, pp. 197–201.
12. King, R.W.; Pasad, S. *Fundamental Electromagnetic Theory and Applications*; Prentice Hall: Upper Saddle River, NJ, 1986, pp. 40–48.
13. Kraus, J.D. *Electromagnetics*; McGraw-Hill: New York, 1953, pp. 148–161, 214–238.
14. Mallison, J.C. *Magneto-Resistive and Spin Valve Heads*; Academic Press: San Diego, 2002.
15. Matson, G.B.; Weiner M.W. Spectroscopy. In *Magnetic Resonance Imaging*; Stark, D.D., Bradley, W.G. Jr. Eds.; Chapter 15, Mosby Year Books; St. Louis, 1992, 438–477.
16. Matveev, A.N. *Electricity and Magnetism*, Mir Publishers, Moscow, 1986.
17. Maxwell, J.C. *A Treatise on Electricity and Magnetism*; Dover Publications: New York, 1954, Vols. 1 and 2.
18. Popović, B.D. *Introductory Engineering Electromagnetics*; Addison-Wesley, Reading, MA, 1971, pp. 243–257, 298–322.
19. Popović Z.; Popović, B.D. *Introductory Electromagnetics*; Prentice Hall: NJ, 1999.
20. Rao, S.; Binder, J.; Bandettini, P.; Hammeke, T.; Yetkin, F.; Jesmanowicz, A.; Lisk, L.; Morris, G.; Mueller, W.; Estkowski, L.; Wong, E.; Haughton, V.; Hyde, J. "Functional magnetic resonance imaging of complex human movements." Neurology **1993**, *43*, 2311–2318.
21. Shen, L.C.; Kong J.A. *Applied Electromagnetism*, 3rd Ed.; PWS Publishing Company: Coston, 1953, pp. 445–462.
22. Slater, J.C.; Frank, H.H. *Electromagnetism*; Dover Publications: New York, 1969, pp. 54–59, 67–70.
23. Silvester, P. *Modern Electromagnetic Fields*; Prentice Hall: Upper Saddle River, NJ, 1968, pp. 172–182.

© 2006 by Taylor & Francis Group, LLC

4
Electromagnetic Induction

Milica Popović
McGill University, Montréal, Quebec

Branko D. Popović[†]
University of Belgrade, Belgrade, Yugoslavia

Zoya Popović
University of Colorado, Boulder, Colorado

To the loving memory of our father, professor, and coauthor. We hope that he would have agreed with the changes we have made after his last edits.

— *Milica and Zoya Popović*

4.1. INTRODUCTION

In 1831 Michael Faraday performed experiments to check whether current is produced in a closed wire loop placed near a magnet, in analogy to dc currents producing magnetic fields. His experiment showed that this could not be done, but Faraday realized that a *time-varying current in the loop was obtained while the magnet was being moved toward it or away from it*. The law he formulated is known as *Faraday's law of electromagnetic induction*. It is perhaps the most important law of electromagnetism. Without it there would be no electricity from rotating generators, no telephone, no radio and television, no magnetic memories, to mention but a few applications.

The phenomenon of electromagnetic induction has a simple physical interpretation. Two charged particles ("charges") at rest act on each other with a force given by Coulomb's law. Two charges moving with *uniform velocities* act on each other with an additional force, the magnetic force. If a particle is *accelerated*, there is another additional force that it exerts on other charged particles, stationary or moving. As in the case of the magnetic force, if only a pair of charges is considered, this additional force is much smaller than Coulomb's force. However, time-varying currents in conductors involve a vast number of accelerated charges, and produce effects significant enough to be easily measurable.

This additional force is of the same *form* as the electric force ($\mathbf{F} = Q\mathbf{E}$). However, other properties of the electric field vector, $\mathbf{E}$ in this case, are different from those of the

[†]Deceased.

© 2006 by Taylor & Francis Group, LLC

electric field vector of static charges. When we wish to stress this difference, we use a slightly different name: the *induced electric field strength*.

The induced electric field and electromagnetic induction have immense practical consequences. Some examples include:

The electric field of electromagnetic waves (e.g., radio waves or light) is basically the induced electric field;
In electrical transformers, the induced electric field is responsible for obtaining higher or lower voltage than the input voltage;
The skin effect in conductors with ac currents is due to induced electric field;
Electromagnetic induction is also the cause of "magnetic coupling" that may result in undesired interference between wires (or metal traces) in any system with time-varying current, an effect that increases with frequency.

The goal of this chapter is to present:

Fundamental theoretical foundations for electromagnetic induction, most importantly Faraday's law;
Important consequences of electromagnetic induction, such as Lentz's law and the skin effect;
Some simple and commonly encountered examples, such as calculation of the inductance of a solenoid and coaxial cable;
A few common applications, such as generators, transformers, electromagnets, etc.

4.2. THEORETICAL BACKGROUND AND FUNDAMENTAL EQUATIONS

4.2.1. The Induced Electric Field

The practical sources of the induced electric field are time-varying currents in a broader sense. If we have, for example, a stationary and rigid wire loop with a time-varying current, it produces an induced electric field. However, a wire loop that changes shape and/or is moving, carrying a *time-constant* current, also produces a time-varying current in space and therefore induces an electric field. Currents equivalent to Ampère's currents in a moving magnet have the same effect and therefore also produce an induced electric field.

Note that in both of these cases there exists, in addition, a time-varying magnetic field. Consequently, a time-varying (induced) electric field is always accompanied by a time-varying magnetic field, and conversely, a time-varying magnetic field is always accompanied by a time-varying (induced) electric field.

The basic property of the induced electric field $\mathbf{E}_{\text{ind}}$ is the same as that of the static electric field: it acts with a force $\mathbf{F} = Q\mathbf{E}_{\text{ind}}$ on a point charge Q. However, the two components of the electric field differ in the work done by the field in moving a point charge around a closed contour. For the static electric field this work is always zero, but for the induced electric field it is not. Precisely this property of the induced electric field gives rise to a very wide range of consequences and applications. Of course, a charge can be situated simultaneously in both a static (Coulomb-type) and an induced field, thus being subjected to a total force

$$\mathbf{F} = Q(\mathbf{E}_{\text{st}} + \mathbf{E}_{\text{ind}}) \tag{4.1}$$

© 2006 by Taylor & Francis Group, LLC

We know how to calculate the static electric field of a given distribution of charges, but how can we determine the induced electric field strength? When a charged particle is moving with a velocity **v** with respect to the source of the magnetic field, the answer follows from the magnetic force on the charge:

$$\mathbf{E}_{\text{ind}} = \mathbf{v} \times \mathbf{B} \qquad (\text{V/m}) \tag{4.2}$$

If we have a current distribution of density **J** (a slowly time-varying function of position) in vacuum, localized inside a volume **v**, the induced electric field is found to be

$$\mathbf{E}_{\text{ind}} = -\frac{\partial}{\partial t}\left(\frac{\mu_0}{4\pi}\int_V \frac{\mathbf{J}\cdot dV}{r}\right) \qquad (\text{V/m}) \tag{4.3}$$

In this equation, r is the distance of the point where the induced electric field is being determined from the volume element dV. In the case of currents over surfaces, $\mathbf{J}(t)\cdot dV$ in Eq. (4.3) should be replaced by $\mathbf{J}_s(t)\cdot dS$, and in the case of a thin wire by $i(t)\cdot dl$.

If we know the distribution of time-varying currents, Eq. (4.3) enables the determination of the induced electric field at any point of interest. Most often it is not possible to obtain the induced electric field strength in analytical form, but it can always be evaluated numerically.

4.2.2. Faraday's Law of Electromagnetic Induction

Faraday's law is an equation for the *total* electromotive force (*emf*) induced in a closed loop due to the induced electric field. This electromotive force is distributed along the loop (not concentrated at a single point of the loop), but we are rarely interested in this distribution. Thus, Faraday's law gives us what is relevant only from the circuit-theory point of view—the *emf* of the Thevenin generator equivalent to all the elemental generators acting in the loop.

Consider a closed conductive contour C, either moving arbitrarily in a time-constant magnetic field or stationary with respect to a system of time-varying currents producing an induced electric field. If the wire segments are moving in a magnetic field, there is an induced field acting along them of the form in Eq. (4.2), and if stationary, the induced electric field is given in Eq. (4.3). In both cases, a segment of the wire loop behaves as an elemental generator of an *emf*

$$de = \mathbf{E}_{\text{ind}}\cdot d\mathbf{l} \tag{4.4}$$

so that the *emf* induced in the entire contour is given by

$$e = \oint_C \mathbf{E}_{\text{ind}}\cdot d\mathbf{l} \tag{4.5}$$

If the *emf* is due to the contour motion only, this becomes

$$e = \oint_C \mathbf{v}\times\mathbf{B}\cdot d\mathbf{l} \tag{4.6}$$

© 2006 by Taylor & Francis Group, LLC

It can be shown that, whatever the cause of the induced electric field (the contour motion, time-varying currents, or the combination of the two), the total *emf* induced in the contour can be expressed in terms of time variation of the magnetic flux through the contour:

$$e = \oint_C \mathbf{E}_{\text{ind}} \cdot d\mathbf{l} = -\frac{d\Phi_{\text{through } C \text{ in } dt}}{dt} = -\frac{d}{dt}\int_S \mathbf{B} \cdot d\mathbf{S} \quad \text{(V)} \tag{4.7}$$

This is *Faraday's law of electromagnetic induction*. The reference direction along the contour, by convention, is connected with the reference direction of the normal to the surface S spanning the contour by the right-hand rule. Note again that the induced *emf* in this equation is nothing but the voltage of the Thévenin generator equivalent to all the elemental generators of electromotive forces $\mathbf{E}_{\text{ind}} \cdot d\mathbf{l}$ acting around the loop. The possibility of expressing the induced *emf* in terms of the magnetic flux alone is not surprising. We know that the induced electric field is always accompanied by a magnetic field, and the above equation only reflects the relationship that exists between the two fields (although the relationship itself is not seen from the equation). Finally, this equation is valid only if the time variation of the magnetic flux through the contour is due either to motion of the contour in the magnetic field or to time variation of the magnetic field in which the contour is situated (or a combination of the two). No other cause of time variation of the magnetic flux will result in an induced *emf*.

4.2.3. Potential Difference and Voltage in a Time-varying Electric and Magnetic Field

The voltage between two points is defined as the line integral of the *total* electric field strength, given in Eq. (4.1), from one point to the other. In electrostatics, the induced electric field does not exist, and voltage does not depend on the path between these points. This is *not* the case in a time-varying electric and magnetic field.

Consider arbitrary time-varying currents and charges producing a time-varying electric and magnetic field, Fig. 4.1. Consider two points, A and B, in this field, and two paths, a and b, between them, as indicated in the figure. The voltage between these two points along the two paths is given by

$$V_{AB\,\text{along}\,a\,\text{or}\,b} = \int_{A\,\text{along}\,a\,\text{or}\,b}^{B} (\mathbf{E}_{\text{st}} + \mathbf{E}_{\text{ind}}) \cdot d\mathbf{l} \tag{4.8}$$

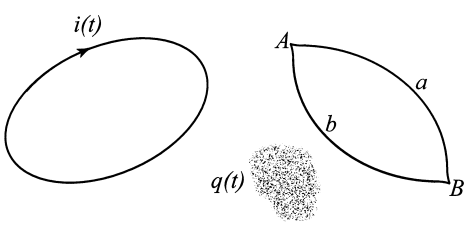


Figure 4.1 An arbitrary distribution of time-varying currents and charges.

© 2006 by Taylor & Francis Group, LLC

The integral between A and B of the static part is simply the potential difference between A and B, and therefore

$$V_{AB\,\text{along}\,a\,\text{or}\,b} = V_A - V_B + \int_{A\,\text{along}\,a\,\text{or}\,b}^{B} \mathbf{E}_{\text{ind}} \cdot d\mathbf{l} \tag{4.9}$$

The potential difference $V_A - V_B$ does not depend on the path between A and B, but the integral in this equation is different for paths a and b. These paths form a closed contour. Applying Faraday's law to that contour, we have

$$e_{\text{induced in closed contour}\,AaBbA} = \oint_{AaBbA} \mathbf{E}_{\text{ind}} \cdot d\mathbf{l} = \int_{AaB} \mathbf{E}_{\text{ind}} \cdot d\mathbf{l} + \int_{BbA} \mathbf{E}_{\text{ind}} \cdot d\mathbf{l} = -\frac{d\Phi}{dt} \tag{4.10}$$

where Φ is the magnetic flux through the surface spanned by the contour $AaBbA$. Since the right side of this equation is generally nonzero, the line integrals of $\mathbf{E}_{\text{ind}}$ from A to B along a and along b are different. Consequently, *the voltage between two points in a time-varying electric and magnetic field depends on the choice of integration path between these two points*.

This is a very important practical conclusion for time-varying electrical circuits. It implies that, contrary to circuit theory, the voltage measured across a circuit by a voltmeter depends on the shape of the leads connected to the voltmeter terminals. Since the measured voltage depends on the rate of change of magnetic flux through the surface defined by the voltmeter leads and the circuit, this effect is particularly pronounced at high frequencies.

4.2.4. Self-inductance and Mutual Inductance

A time-varying current in one current loop induces an *emf* in another loop. In linear media, an electromagnetic parameter that enables simple determination of this *emf* is the *mutual inductance*.

A wire loop with time-varying current creates a time-varying induced electric field not only in the space around it but also along the loop itself. As a consequence, there is a feedback—the current produces an effect which affects itself. The parameter known as *inductance*, or *self-inductance*, of the loop enables simple evaluation of this effect.

Consider two stationary thin conductive contours C_1 and C_2 in a linear medium (e.g., air), shown in Fig. 4.2. When a time-varying current $i_1(t)$ flows through the first contour, it creates a time-varying magnetic field, as well as a time-varying induced electric field, $\mathbf{E}_{1\,\text{ind}}(t)$. The latter produces an *emf* $e_{12}(t)$ in the second contour, given by

$$e_{12}(t) = \oint_{C_2} \mathbf{E}_{1\,\text{ind}} \cdot d\mathbf{l}_2 \tag{4.11}$$

where the first index denotes the source of the field (contour 1 in this case).

It is usually much easier to find the induced *emf* using Faraday's law than in any other way. The magnetic flux density vector in linear media is proportional to the current

© 2006 by Taylor & Francis Group, LLC

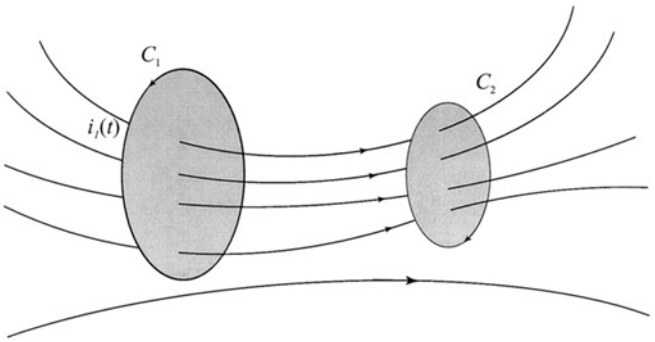


Figure 4.2 Two coupled conductive contours.

that causes the magnetic field. It follows that the flux $\Phi_{12}(t)$ through C_2 caused by the current $i_1(t)$ in C_1 is also proportional to $i_1(t)$:

$$\Phi_{12}(t) = L_{12} \cdot i_1(t) \tag{4.12}$$

The proportionality constant L_{12} is the *mutual inductance* between the two contours. This constant depends only on the geometry of the system and the properties of the (linear) medium surrounding the current contours. Mutual inductance is denoted both by L_{12} or sometimes in circuit theory by M.

Since the variation of $i_1(t)$ can be arbitrary, the same expression holds when the current through C_1 is a dc current:

$$\Phi_{12} = L_{12} I_1 \tag{4.13}$$

Although mutual inductance has no practical meaning for dc currents, this definition is used frequently for the determination of mutual inductance.

According to Faraday's law, the *emf* can alternatively be written as

$$e_{12}(t) = -\frac{d\Phi_{12}}{dt} = -L_{12}\frac{di_1(t)}{dt} \tag{4.14}$$

The unit for inductance, equal to a Wb/A, is called a *henry* (H). One henry is quite a large unit. Most frequent values of mutual inductance are on the order of a mH, μH, or nH.

If we now assume that a current $i_2(t)$ in C_2 causes an induced *emf* in C_1, we talk about a mutual inductance L_{21}. It turns out that $L_{12} = L_{21}$ always. [This follows from the expression for the induced electric field in Eqs. (4.3) and (4.5).] So, we can write

$$L_{12} = \frac{\Phi_{12}}{I_1} = L_{21} = \frac{\Phi_{21}}{I_2} \qquad \text{(H)} \tag{4.15}$$

These equations show that we need to calculate either Φ_{12} or Φ_{21} to determine the mutual inductance, which is a useful result since in some instances one of these is much simpler to calculate than the other.

© 2006 by Taylor & Francis Group, LLC

Note that mutual inductance can be negative as well as positive. The sign depends on the actual geometry of the system and the adopted reference directions along the two loops: if the current in the reference direction of one loop produces a positive flux in the other loop, then mutual inductance is positive, and vice versa. For calculating the flux, the normal to the loop surface is determined by the right-hand rule with respect to its reference direction.

As mentioned, when a current in a contour varies in time, the induced electric field exists everywhere around it and therefore also along its entire length. Consequently, there is an induced *emf* in the contour itself. This process is known as *self-induction*. The simplest (even if possibly not physically the clearest) way of expressing this *emf* is to use Faraday's law:

$$e(t) = -\frac{d\Phi_{\text{self}}(t)}{dt} \tag{4.16}$$

If the contour is in a linear medium (i.e., the flux through the contour is proportional to the current), we define the *self-inductance* of the contour as the ratio of the flux $\Phi_{\text{self}}(t)$ through the contour due to current $i(t)$ in it and $i(t)$,

$$L = \frac{\Phi_{\text{self}}(t)}{i(t)} \qquad \text{(H)} \tag{4.17}$$

Using this definition, the induced *emf* can be written as

$$e(t) = -L\frac{di(t)}{dt} \tag{4.18}$$

The constant L depends only on the geometry of the system, and its unit is again a henry (H). In the case of a dc current, $L = \Phi/I$, which can be used for determining the self-inductance in some cases in a simple manner.

The self-inductances of two contours and their mutual inductance satisfy the following condition:

$$L_{11}L_{22} \geq L_{12}^2 \tag{4.19}$$

Therefore, the largest possible value of mutual inductance is the geometric mean of the self-inductances. Frequently, Eq. (4.19) is written as

$$L_{12} = k\sqrt{L_{11}L_{22}} \qquad -1 \leq k \leq 1 \tag{4.20}$$

The dimensionless coefficient k is called the *coupling coefficient*.

4.2.5. Energy and Forces in the Magnetic Field

There are many devices that make use of electric or magnetic forces. Although this is not commonly thought of, almost any such device can be made in an "electric version" and in a "magnetic version." We shall see that the magnetic forces are several orders of magnitude stronger than electric forces. Consequently, devices based on magnetic forces

© 2006 by Taylor & Francis Group, LLC

are much smaller in size, and are used more often when force is required. For example, electric motors in your household and in industry, large cranes for lifting ferromagnetic objects, home bells, electromagnetic relays, etc., all use magnetic, not electric, forces.

A powerful method for determining magnetic forces is based on energy contained in the magnetic field. While establishing a dc current, the current through a contour has to change from zero to its final dc value. During this process, there is a changing magnetic flux through the contour due to the changing current, and an *emf* is induced in the contour. This *emf* opposes the change of flux (see Lentz's law in Sec. 4.3.2). In order to establish the final static magnetic field, the sources have to overcome this *emf*, i.e., to spend some energy. A part (or all) of this energy is stored in the magnetic field and is known as *magnetic energy*.

Let n contours, with currents $i_1(t), i_2(t), \ldots, i_n(t)$ be the sources of a magnetic field. Assume that the contours are connected to generators of electromotive forces $e_1(t), e_2(t), \ldots, e_n(t)$. Finally, let the contours be stationary and rigid (i.e., they cannot be deformed), with total fluxes $\Phi_1(t), \Phi_2(t), \ldots, \Phi_n(t)$. If the medium is linear, energy contained in the magnetic field of such currents is

$$W_m = \frac{1}{2}\sum_{k=1}^{n} I_k \Phi_k \tag{4.21}$$

This can be expressed also in terms of self- and mutual inductances of the contours and the currents in them, as

$$W_m = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} L_{jk} I_j I_k \tag{4.22}$$

which for the important case of a single contour becomes

$$W_m = \frac{1}{2} I\Phi = \frac{1}{2} LI^2 \tag{4.23}$$

If the medium is ferromagnetic these expressions are not valid, because at least one part of the energy used to produce the field is transformed into heat. Therefore, for ferromagnetic media it is possible only to evaluate the total energy used to obtain the field. If B_1 is the initial magnetic flux density and B_2 the final flux density at a point, energy density spent in order to change the magnetic flux density vector from B_1 to B_2 at that point is found to be

$$\frac{dA_m}{dV} = \int_{B_1}^{B_2} H(t) \cdot dB(t) \qquad (\mathrm{J/m^3}) \tag{4.24}$$

In the case of linear media (see Chapter 3), energy used for changing the magnetic field is stored in the field, i.e., $dA_m = dW_m$. Assuming that the B field changed from zero to some value B, the volume density of magnetic energy is given by

$$\frac{dW_m}{dV} = \int_b^B \frac{B}{\mu} \cdot dB = \frac{1}{2}\frac{B^2}{\mu} = \frac{1}{2}\mu H^2 = \frac{1}{2} BH \qquad (\mathrm{J/m^3}) \tag{4.25}$$

© 2006 by Taylor & Francis Group, LLC

The energy in a *linear medium* can now be found by integrating this expression over the entire volume of the field:

$$W_m = \int_V \frac{1}{2}\mu H^2 \, dV \qquad \text{(J)} \tag{4.26}$$

If we know the distribution of currents in a magnetically homogeneous medium, the magnetic flux density is obtained from the Biot-Savart law. Combined with the relation $d\mathbf{F}_m = I \cdot d\mathbf{l} \times \mathbf{B}$, we can find the magnetic force on any part of the current distribution. In many cases, however, this is quite complicated.

The magnetic force can also be evaluated as a derivative of the magnetic energy. This can be done assuming either (1) the fluxes through all the contours are kept constant or (2) the currents in all the contours are kept constant. In some instances this enables very simple evaluation of magnetic forces.

Assume first that during a displacement dx of a body in the magnetic field along the x axis, we keep the fluxes through all the contours constant. This can be done by varying the currents in the contours appropriately. The x component of the magnetic force acting on the body is then obtained as

$$F_x = -\left(\frac{dW_m}{dx}\right)_{\Phi = \text{const}} \tag{4.27}$$

In the second case, when the currents are kept constant,

$$F_x = +\left(\frac{dW_m}{dx}\right)_{I = \text{const}} \tag{4.28}$$

The signs in the two expressions for the force determine the direction of the force. In Eq. (4.28), the positive sign means that when current sources are producing all the currents in the system ($I = \text{const}$), the magnetic field energy increases, as the generators are the ones that add energy to the system and produce the force.

4.3. CONSEQUENCES OF ELECTROMAGNETIC INDUCTION

4.3.1. Magnetic Coupling

Let a time-varying current $i(t)$ exist in a circular loop C_1 of radius a, Fig. 4.3. According to Eq. (4.3), lines of the induced electric field around the loop are circles centered at the loop axis normal to it, so that the line integral of the induced electric field around a circular contour C_2 indicated in the figure in dashed line *is not zero*. If the contour C_2 is a wire loop, this field acts as a distributed generator along the entire loop length, and a current is induced in that loop.

The reasoning above does not change if loop C_2 is not circular. We have thus reached an extremely important conclusion: *The induced electric field of time-varying currents in one wire loop produces a time-varying current in an adjacent closed wire loop*. Note that the other loop need not (and usually does not) have any physical contact with the first loop. This means that the induced electric field enables transport of energy from one loop to the other through vacuum. Although this coupling is actually obtained by means of the induced electric field, it is known as *magnetic coupling*.

© 2006 by Taylor & Francis Group, LLC

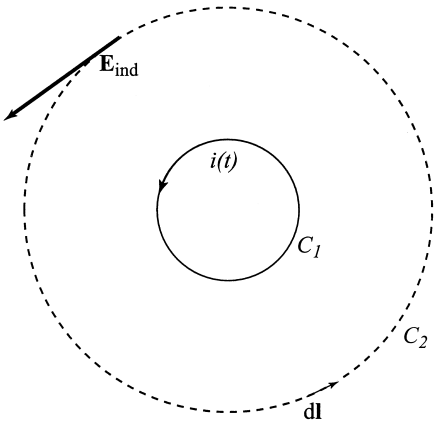


Figure 4.3 A circular loop C_1 with a time-varying current $i(t)$. The induced electric field of this current is tangential to the circular loop C_2 indicated in dashed line, so that it results in a distributed emf around the loop.

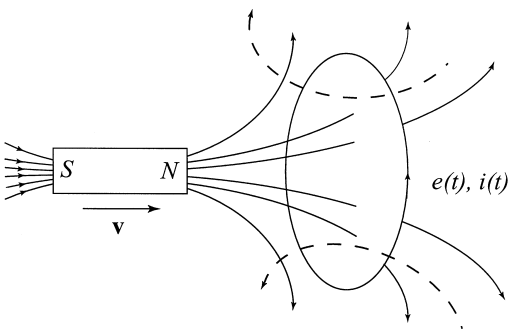


Figure 4.4 Illustration of Lentz's law.

Note that if the wire loop C_2 is not closed, the induced field nevertheless induces distributed generators along it. The loop behaves as an open-circuited equivalent (Thévenin) generator.

4.3.2. Lentz's Law

Figure 4.4 shows a permanent magnet approaching a stationary loop. The permanent magnet is equivalent to a system of macroscopic currents. Since it is moving, the magnetic flux created by these currents through the contour varies in time. According to the reference direction of the contour shown in the figure, the change of flux is positive, $(d\Phi/dt) > 0$, so the induced *emf* is in the direction shown in the figure. The *emf* produces a current through the closed loop, which in turn produces its own magnetic field, shown in the figure in dashed line. As a result, the change of the magnetic flux, caused initially by the magnet motion, is reduced. This is *Lentz's law*: the induced current in a conductive contour tends to decrease the *change* in magnetic flux through the contour. Lentz's law describes a feedback property of electromagnetic induction.

© 2006 by Taylor & Francis Group, LLC

4.3.3. Eddy Currents

A very important consequence of the induced electric field are eddy currents. These are currents induced throughout a solid metal body when the body is situated in a time-varying magnetic (i.e., induced electric) field.

As the first consequence of eddy currents, there is power lost to heat according to Joule's law. Since the magnitude of eddy currents is proportional to the magnitude of the induced electric field, eddy-current losses are proportional to the square of frequency.

As the second consequence, there is a secondary magnetic field due to the induced currents which, following Lentz's law, reduces the magnetic field inside the body. Both of these effects are usually not desirable. For example, in a ferromagnetic core shown in Fig. 4.5, Lentz's law tells us that eddy currents tend to decrease the flux in the core, and the magnetic circuit of the core will not be used efficiently. The flux density vector is the smallest at the center of the core, because there the **B** field of all the induced currents adds up. The total magnetic field distribution in the core is thus nonuniform.

To reduce these two undesirable effects, ferromagnetic cores are made of mutually insulated thin sheets, as shown in Fig. 4.6. Now the flux through the sheets is encircled by much smaller loops, the *emf* induced in these loops is consequently much smaller, and so the eddy currents are also reduced significantly. Of course, this only works if vector **B** is parallel to the sheets.

In some instances, eddy currents are created on purpose. For example, in induction furnaces for melting metals, eddy currents are used to heat solid metal pieces to melting temperatures.

4.3.4. The Skin Effect and the Proximity and Edge Effects

A time-invariant current in a homogeneous cylindrical conductor is distributed uniformly over the conductor cross section. If the conductor is not cylindrical, the time-invariant current in it is not distributed uniformly, *but it exists in the entire conductor*. A time-varying current has a tendency to concentrate near the surfaces of conductors. At very high frequencies, the current is restricted to a very thin layer near the conductor surface, practically on the surfaces themselves. Because of this extreme case, the entire phenomenon of nonuniform distribution of time-varying currents in conductors is known as the *skin effect*.

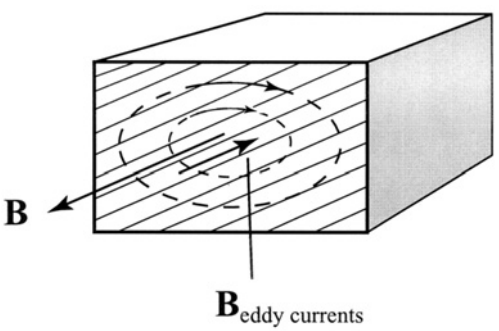
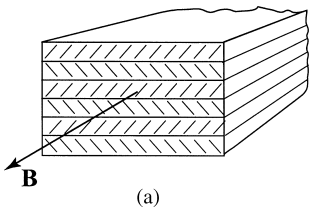


Figure 4.5 Eddy currents in a piece of ferromagnetic core. Note that the total **B** field in the core is reduced due to the opposite field created by eddy currents.

© 2006 by Taylor & Francis Group, LLC



(a)

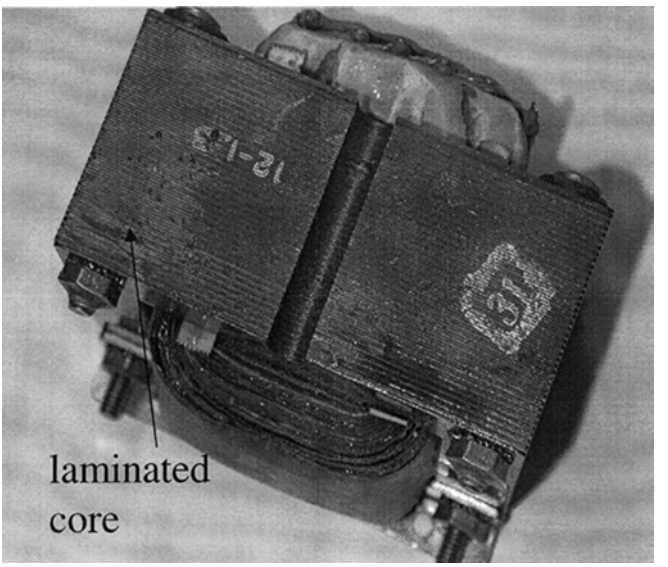


Figure 4.6 A ferromagnetic core for transformers and ac machines consists of thin insulated sheets: (a) sketch of core and (b) photograph of a typical transformer core.

The cause of skin effect is electromagnetic induction. A time-varying magnetic field is accompanied by a time-varying induced electric field, which in turn creates secondary time-varying currents (induced currents) and a secondary magnetic field. The induced currents produce a magnetic flux which opposes the external flux (the same flux that "produced" the induced currents). As a consequence, the total flux is reduced. The larger the conductivity, the larger the induced currents are, and the larger the permeability, the more pronounced the flux reduction is. Consequently, both the total time-varying magnetic field and induced currents inside conductors are reduced when compared with the dc case.

The skin effect is of considerable practical importance. For example, at very high frequencies a very thin layer of conductor carries most of the current. Any conductor (or for that matter, any other material), can be coated with silver (the best available conductor) and practically the entire current will flow through this thin silver coating. Even at power frequencies in the case of high currents, the use of thick solid conductors is not efficient, and bundled conductors are used instead.

The skin effect exists in all conductors, but, as mentioned, the tendency of current and magnetic flux to be restricted to a thin layer on the conductor surface is much more pronounced for a ferromagnetic conductor than for a nonferromagnetic conductor of the same conductivity. For example, for iron at 60 Hz the thickness of this layer is on the order

© 2006 by Taylor & Francis Group, LLC

of only 0.5 mm. Consequently, solid ferromagnetic cores for alternating current electric motors, generators, transformers, etc., would result in poor use of the ferromagnetic material and high losses. Therefore, laminated cores made of thin, mutually insulated sheets are used instead. At very high frequencies, ferrites (ferrimagnetic ceramic materials) are used, because they have very low conductivity when compared to metallic ferromagnetic materials.

Consider a body with a sinusoidal current of angular frequency ω and let the material of the body have a conductivity σ and permeability μ. If the frequency is high enough, the current will be distributed over a very thin layer over the body surface, the current density being maximal at the surface (and parallel to it), and decreasing rapidly with the distance z from it:

$$J(z) = J_0 e^{-jz/\delta} \tag{4.29}$$

where

$$\delta = \sqrt{\frac{\omega\mu\sigma}{2}} = \sqrt{\pi\mu\sigma}\sqrt{f} \tag{4.30}$$

The intensity of the current density vector decreases exponentially with increasing z. At a distance δ the amplitude of the current density vector decreases to $1/e$ of its value J_0 at the boundary surface. This distance is known as the *skin depth*. For example, for copper ($\sigma = 57 \times 10^6$ S/m, $\mu = \mu_0$), the skin depth at 1 MHz is only 0.067 mm. For iron ($\sigma = 10^7$ S/m, $\mu_r = 1000$), the skin depth at 60 Hz is 0.65 mm, and for sea water ($\sigma = 4$ S/m, $\mu = \mu_0$), at the same frequency it is 32.5 m. Table 4.1 summarizes the value of skin depth in some common materials at a few characteristic frequencies.

The result for skin depth for iron at power frequencies (50 Hz or 60 Hz), $\delta \cong 5$ mm, tells us something important. Iron has a conductivity that is only about six times less than that of copper. On the other hand, copper is much more expensive than iron. Why do we then not use iron wires for the distribution of electric power in our homes? Noting that there are millions of kilometers of such wires, the savings would be very large. Unfortunately, due to a large relative permeability—iron has very small power-frequency skin depth (a fraction of a millimeter)—the losses in iron wire are large, outweighing the savings, so copper or aluminum are used instead.

Keeping the current intensity the same, Joule losses increase with frequency due to increased resistance in conductors resulting from the skin effect. It can be shown that Joule's losses per unit area are given by

$$\frac{dP_J}{dS} = R_s|H_0|^2 \qquad (\mathrm{W/m^2}) \tag{4.31}$$

Table 4.1 Values of Skin Depth for Some Common Materials at 60 Hz, 1 kHz, 1 MHz, and 1 GHz.

Material	$f = 60$ Hz	$f = 1$ kHz	$f = 1$ MHz	$f = 1$ GHz
Copper	8.61 mm	2.1 mm	0.067 mm	2.11 μm
Iron	0.65 mm	0.16 mm	5.03 μm	0.016 μm
Sea water	32.5 m	7.96 m	0.25 m	7.96 mm
Wet soil	650 m	159m	5.03 m	0.16 m

© 2006 by Taylor & Francis Group, LLC

where H_0 is the complex rms value of the tangential component of the vector **H** on the conductor surface, and R_s is the *surface resistance* of the conductor, given by

$$R_s = \sqrt{\frac{\omega\mu}{2\sigma}} \qquad (\Omega) \tag{4.32}$$

Equation (4.32) is used for determining the attenuation in all metal waveguides, such as two-wire lines (twin-lead), coaxial lines, and rectangular waveguides.

The term *proximity effect* refers to the influence of alternating current in one conductor on the current distribution in another nearby conductor. Consider a coaxial cable of finite length. Assume for the moment that there is an alternating current only in the inner conductor (for example, that it is connected to a generator), and that the outer conductor is not connected to anything. If the outer conductor is much thicker than the skin depth, there is practically no magnetic field inside the outer conductor. If we apply Ampère's law to a coaxial circular contour contained in that conductor, it follows that the induced current on the *inside* surface of the outer conductor is exactly equal and opposite to the current in the inner conductor. This is an example of the proximity effect. If in addition there is normal cable current in the outer conductor, it is the same but opposite to the current on the conductor outer surface, so the two cancel out. We are left with a current over the inner conductor and a current over the inside surface of the outer conductor. This combination of the skin and proximity effects is what is usually actually encountered in practice.

Redistribution of Current on Parallel Wires and Printed Traces

Consider as the next example three long parallel wires a certain distance apart lying in one plane. The three ends are connected together at one and at the other end of the wires, and these common ends are connected by a large loop to a generator of sinusoidal *emf*. Are the currents in the three wires the same? At first glance we should expect them to be the same, but due to the induced electric field they are not: the current intensity in the middle wire will always be smaller than in the other two.

The above example is useful for understanding the distribution of ac current across the cross section of a printed metal strip, such as a trace on a printed-circuit board. The distribution of current across the strip will not be uniform (which it is at zero frequency). The current amplitude will be much greater along the strip edges than along its center. This effect is sometimes referred to as the *edge effect*, but it is, in fact, the skin effect in strip conductors. Note that for a strip line (consisting of two close parallel strips) this effect is very small because the induced electric fields due to opposite currents in the two strips practically cancel out.

4.3.5. Limitations of Circuit Theory

Circuit theory is the basic tool of electrical engineers, but it is approximate and therefore has limitations. These limitations can be understood only using electromagnetic-field theory. We consider here the approximations implicit in Kirchhoff's voltage law (KVL). This law states that the sum of voltages across circuit branches along any closed path is zero and that voltages and currents in circuit branches do not depend on the circuit actual geometrical shape. Basically, this means that this law neglects the induced electric field

© 2006 by Taylor & Francis Group, LLC

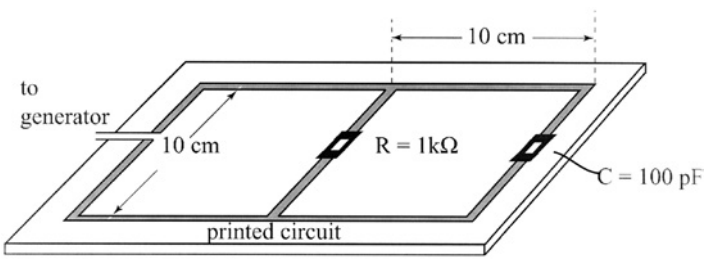
produced by currents in the circuit branches. This field increases with frequency, so that at a certain frequency (depending on circuit properties and its actual size) the influence of the induced electric field on circuit behavior becomes of the same order of magnitude as that due to generators in the circuit. The analysis of circuit behavior in such cases needs to be performed by electromagnetic analysis, usually requiring numerical solutions.

As a simple example, consider the circuit in Fig. 4.7, consisting of several printed traces and two lumped (pointlike, or much smaller than a wavelength) surface-mount components. For a simple two-loop circuit 10 cm × 20 cm in size, already at a frequency of 10 MHz circuit analysis gives results with errors exceeding 20%. The tabulated values in Fig. 4.7 show the calculated and measured complex impedance seen by the generator at different frequencies.

Several useful practical conclusions can be drawn. The first is that for circuits that contain wires or traces and low-valued resistors, this effect will become pronounced at lower frequencies. The second is that the behavior of an ac circuit always depends on the circuit shape, although in some cases this effect might be negligible. (A complete electromagnetic numerical solution of this circuit would give exact agreement with theory.) This directly applies to measurements of ac voltages (and currents), since the leads of the meter are also a part of the circuit. Sometimes, there is an *emf* induced in the meter leads due to flux through loops formed by parts of the circuit and the leads. This can lead to errors in voltage measurements, and the loops that give rise to the error *emf* are often referred to as *ground loops*.

4.3.6. Superconducting Loops

Some substances have zero resistivity at very low temperatures. For example, lead has zero resistivity below about 7.3 K (just a little bit warmer than liquid helium). This phenomenon is known as *superconductivity*, and such conductors are said to be



Frequency	Calculated Re(Z)	Measured Re(Z)	Calculated Im(Z)	Measured Im(Z)
10 MHz	25 Ω	20 Ω	−150 Ω	−110 Ω
20 MHz	6 Ω	1 Ω	−90 Ω	≈0 Ω
50 MHz	1 Ω	5 Ω	−50 Ω	+180 Ω
100 MHz	≈0 Ω	56 Ω	−15 Ω	+470 Ω

Figure 4.7 Example of impedance seen by the generator for a printed circuit with a surface-mount resistor and capacitor. The table shows a comparison of results obtained by circuit theory and measured values, indicating the range of validity of circuit theory.

© 2006 by Taylor & Francis Group, LLC

superconductors. Some ceramic materials (e.g., yttrium barium oxide) become superconductors at temperatures as "high" as about 70 K (corresponding to the temperature of liquid nitrogen). Superconducting loops have an interesting property when placed in a time-varying magnetic field. The Kirchhoff voltage law for such a loop has the form

$$-\frac{d\Phi}{dt} = 0 \tag{4.33}$$

since the *emf* in the loop is $-d\Phi/dt$ and the loop has zero resistance. From this equation, it is seen that the flux through a superconducting loop remains *constant*. Thus, it is not possible to change the magnetic flux through a superconducting loop by means of electromagnetic induction. The physical meaning of this behavior is the following: If a superconducting loop is situated in a time-varying induced electric field, the current induced in the loop must vary in time so as to produce exactly the same induced electric field in the loop, but in the opposite direction. If this were not so, infinite current would result.

4.4. APPLICATIONS OF ELECTROMAGNETIC INDUCTION AND FARADAY'S LAW

4.4.1. An AC Generator

An ac generator, such as the one sketched in Fig. 4.8, can be explained using Faraday's law. A rectangular wire loop is rotating in a uniform magnetic field (for example, between the poles of a magnet). We can measure the induced voltage in the wire by connecting a voltmeter between contacts C_1 and C_2. Vector **B** is perpendicular to the contour axis. The loop is rotating about this axis with an angular velocity ω. If we assume that at $t = 0$ vector **B** is parallel to vector **n** normal to the surface of the loop, the induced *emf* in the loop is given by

$$e(t) = -\frac{d\Phi(t)}{dt} = \omega abB \sin \omega t = E_{\max} \sin \omega t \tag{4.34}$$

In practice, the coil has many turns of wire instead of a single loop, to obtain a larger induced *emf*. Also, usually the coil is not rotating, but instead the magnetic field is rotating around it, which avoids sliding contacts of the generator.

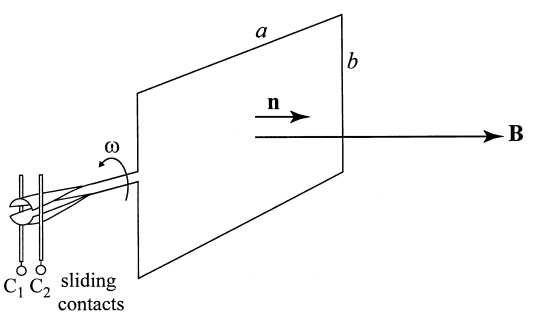


Figure 4.8 A simple ac generator.

© 2006 by Taylor & Francis Group, LLC

4.4.2. Induction Motors

Motors transform electric to mechanical power through interaction of magnetic flux and electric current [1,20,26]. Electric motors are broadly categorized as ac and dc motors, with a number of subclassifications in each category. This section describes the basic operation of induction motors, which are most often encountered in industrial use.

The principles of the polyphase induction motor are here explained on the example of the most commonly used three-phase version. In essence, an induction motor is a transformer. Its magnetic circuit is separated by an air gap into two portions. The fixed *stator* carries the primary winding, and the movable *rotor* the secondary winding, as shown in Fig. 4.9a. An electric power system supplies alternating current to the primary

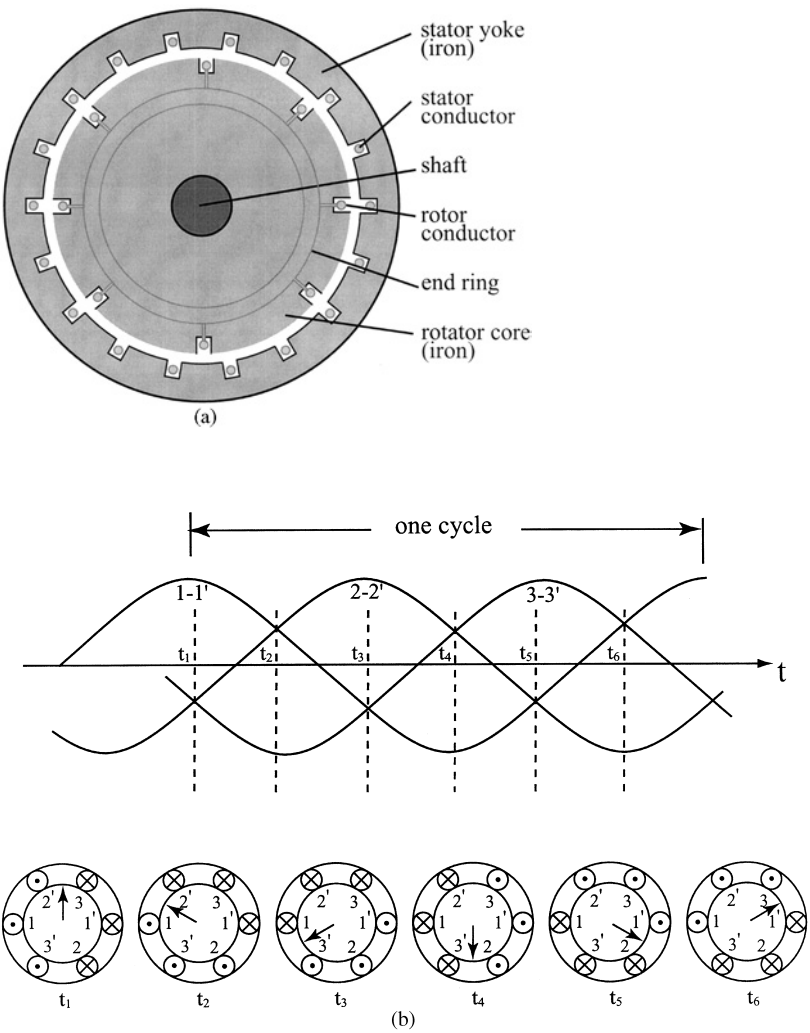


Figure 4.9 (a) Cross section of a three-phase induction motor. 1-1′, 2-2′, and 3-3′ mark the primary stator windings, which are connected to an external three-phase power supply. (b) Time-domain waveforms in the windings of the stator and resulting magnetic field vector rotation as a function of time.

© 2006 by Taylor & Francis Group, LLC

winding, which induces currents in the secondary (short-circuited or closed through an external impedance) and thus causes the motion of the rotor. The key distinguishable feature of this machine with respect to other motors is that the current in the secondary is produced only by electromagnetic induction, i.e., not by an external power source.

The primary windings are supplied by a three-phase system currents, which produce three stationary *alternating* magnetic fields. Their superposition yields a sinusoidally distributed magnetic field in the air gap of the stator, revolving synchronously with the power-supply frequency. The field completes one revolution in one cycle of the stator current, as illustrated in Fig. 4.9b. Thus, the combined effect of three-phase alternating currents with the shown angular arrangement in the stator, results in a *rotating* magnetic field with a constant magnitude and a mechanical angular speed that depends on the frequency of the electric supply.

Two main types of induction motors differ in the configuration of the secondary windings. In squirrel-cage motors, the secondary windings of the rotor are constructed from conductor bars, which are short-circuited by end rings. In the wound-rotor motors, the secondary consists of windings of discrete conductors with the same number of poles as in the primary stator windings.

4.4.3. Electromagnetic Measurement of Fluid Velocity

The velocity of flowing liquids that have a small, but finite, conductivity can be measured using electromagnetic induction. In Fig. 4.10, the liquid is flowing through a flat insulating pipe with an unknown velocity $\mathbf{v}$. The velocity of the fluid is roughly uniform over the cross section of the pipe. To measure the fluid velocity, the pipe is in a magnetic field with a flux density vector $\mathbf{B}$ normal to the pipe. Two small electrodes are in contact with the fluid at the two ends of the pipe cross section. A voltmeter with large input impedance shows a voltage V when connected to the electrodes. The velocity of the fluid is then given by $v = V/B$.

4.4.4. Measurement of AC Currents

A useful application of the induced electric field is for measurement of a sinusoidal current in a conductor without breaking the circuit (as required by standard current measurement). Figure 4.11 shows a conductor with a sinusoidal current of amplitude I_m and angular frequency ω flowing through it. The conductor is encircled by a flexible thin rubber strip of cross-sectional area S, densely wound along its length with N' turns of wire per unit length. We show that if we measure the amplitude of the voltage between the terminals of the strip winding, e.g., V_m, we can calculate I_m.

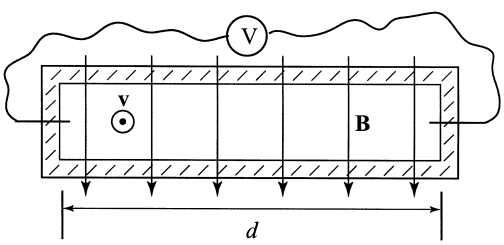


Figure 4.10 Measurement of fluid velocity.

© 2006 by Taylor & Francis Group, LLC

There are $dN = N'dl$ turns of wire on a length dl of the strip. The magnetic flux through a single turn is $\Phi_0 = \mathbf{B} \cdot \mathbf{S}$, and that through dN turns is

$$d\Phi = \Phi_0 \, dN = N'S \, d\mathbf{l} \cdot \mathbf{B} \tag{4.35}$$

The total flux through all the turns of the flexible solenoid is thus

$$\Phi = \oint_C d\Phi = N'S \oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 N'S \cdot i(t) \tag{4.36}$$

according to Ampère's law applied to the contour C along the strip. The induced *emf* in the winding is $e = -d\Phi/dt$, so that, finally, the expression for the amplitude of $i(t)$ reads $I_m = V_m/\mu_0 N'S\omega$.

4.4.5. Problems in Measurement of AC Voltage

As an example of the measurement of ac voltage, consider a straight copper wire of radius $a = 1\,\text{mm}$ with a sinusoidal current $i(t) = 1\cos\omega t\,\text{A}$. A voltmeter is connected between points 1 and 2, with leads as shown in Fig. 4.12. If $b = 50\,\text{cm}$ and $c = 20\,\text{cm}$, we will evaluate the voltage measured by the voltmeter for (1) $\omega = 314\,\text{rad/s}$, (2) $\omega = 10^4\,\text{rad/s}$,

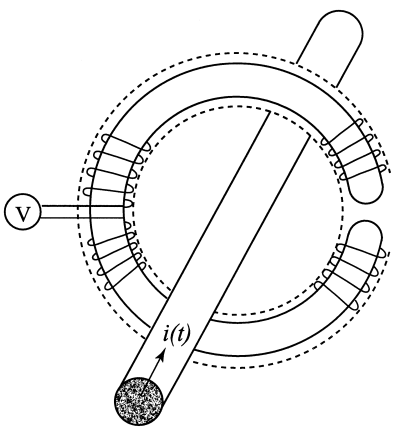


Figure 4.11 A method for measuring ac current in a conductor without inserting an ammeter into the circuit.

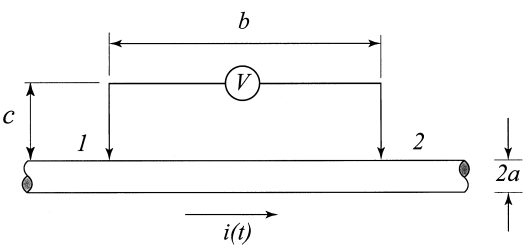


Figure 4.12 Measurement of ac voltage.

© 2006 by Taylor & Francis Group, LLC

and (3) $\omega = 10^6$ rad/s. We assume that the resistance of the copper conductor per unit length, R', is approximately as that for a dc current (which actually is *not* the case, due to skin effect). We will evaluate for the three cases the potential difference $V_1 - V_2 = R'b \cdot i(t)$ and the voltage induced in the leads of the voltmeter.

The voltage measured by the voltmeter (i.e., the voltage between its *very* terminals, and *not* between points 1 and 2) is

$$V_{\text{voltmeter}} = (V_1 - V_2) - e = R'bi - e \tag{4.37}$$

where $R' = 1/\sigma_{\text{Cu}}\pi a^2$, ($\sigma_{\text{Cu}} = 5.7 \times 10^7$ S/m), and e is the induced *emf* in the rectangular contour containing the voltmeter and the wire segment between points 1 and 2 (we neglect the size of the voltmeter). This *emf* is approximately given by

$$e = \frac{\mu_0 b}{2\pi}\frac{di}{dt}\ln\frac{c+a}{a} \tag{4.38}$$

The rms value of the potential difference $(V_1 - V_2)$ amounts to 1.97 mV, and does not depend on frequency. The difference between this potential difference and the voltage indicated by the voltmeter for the three specified frequencies is (1) 117.8 μV, (2) 3.74 μV, and (3) 3.74 V. This difference represents an error in measuring the potential difference using the voltmeter with such leads. We see that in case (2) the relative error is as large as 189%, and that in case (3) such a measurement is meaningless.

4.4.6. Readout of Information Stored on a Magnetic Disk

When a magnetized disk with small permanent magnets (created in the writing process) moves in the vicinity of the air gap of a magnetic head, it will produce time-variable flux in the head magnetic core and the read-and-write coil wound around the core. As a result, an *emf* will be induced in the coil reflecting the magnetization of the disk, in the form of positive and negative pulses. This is sketched in Fig. 4.13. (For the description of the writing process and a sketch of the magnetic head, please see Chapter 3.)

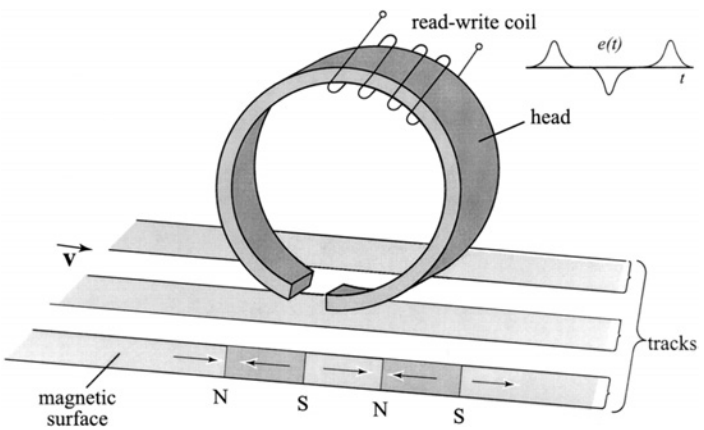


Figure 4.13 A hard disk magnetized through the write process induces a emf in the read process: when the recorded magnetic domains change from south to north pole or vice versa, a voltage pulse proportional to the remanent magnetic flux density is produced. The pulse can be negative or positive.

© 2006 by Taylor & Francis Group, LLC

Historical Note: Magnetic Core Memories

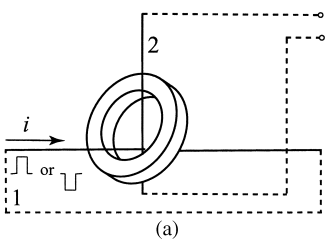
In the readout process of magnetic core memories described in Chapter 3, a negative pulse is passed through circuit 1 in Fig. 4.14. If the core is magnetized to a "1" (positive remanent magnetic flux density of the hysteresis curve), the negative current pulse brings it to the negative tip of the hysteresis loop, and after the pulse is over, the core will remain at the negative remanent flux density point. If, on the other hand, the core is at "0" (negative remanent magnetic flux density of the hysteresis curve), the negative current pulse will make the point go to the negative tip of the hysteresis loop and again end at the point where it started.

While the above described process is occurring, an *emf* is induced in circuit 2, resulting in one of the two possible readings shown in Fig. 4.14. These two pulses correspond to a "1" and a "0." The speed at which this process occurs is about 0.5–5 μs.

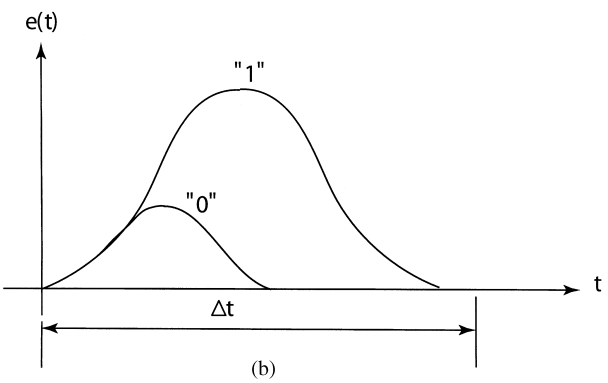
4.4.7. Transformers

A transformer is a magnetic circuit with (usually) two windings, the "primary" and the "secondary," on a common ferromagnetic core, Fig. 4.15. When an ac voltage is applied to the primary coil, the magnetic flux through the core is the same at the secondary and induces a voltage at the open ends of the secondary winding. Ampère's law for this circuit can be written as

$$N_1 i_1 - N_2 i_2 = HL \tag{4.39}$$



(a)



(b)

Figure 4.14 (a) A magnetic core memory bit and (b) induced voltage pulses during the readout process.

© 2006 by Taylor & Francis Group, LLC

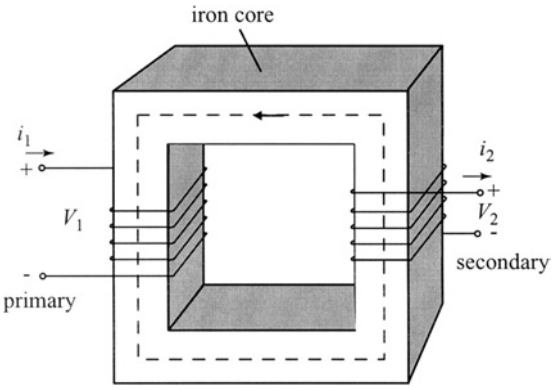


Figure 4.15 Sketch of a transformer with the primary and secondary windings wound on a ferromagnetic core.

where N_1 and N_2 are the numbers of the primary and secondary windings, i_1 and i_2 are the currents in the primary and secondary coils when a generator is connected to the primary and a load to the secondary, H is the magnetic field in the core, and L is the effective length of the core. Since $H = B/\mu$ and, for an *ideal* core, $\mu \to \infty$, both B and H in the ideal core are zero (otherwise the magnetic energy in the core would be infinite). Therefore, for an ideal transformer

$$\frac{i_1}{i_2} = \frac{N_2}{N_1} \tag{4.40}$$

This is the relationship between the primary and secondary currents in an *ideal transformer*. For good ferromagnetic cores, the permeability is high enough that this is a good approximation.

From the definition of magnetic flux, the flux through the core is proportional to the number of windings in the primary. From Faraday's law, the induced *emf* in the secondary is proportional to the number of times the magnetic flux in the core passes through the surface of the secondary windings, i.e., to N_2. (This is even more evident if one keeps in mind that the lines of induced electric field produced by the primary current encircle the core, i.e., going along the secondary winding the integral of the induced electric field is N_2 times that for a single turn.) Therefore, the following can be written for the voltages across the primary and secondary windings:

$$\frac{v_1}{v_2} = \frac{N_1}{N_2} \tag{4.41}$$

Assume that the secondary winding of an ideal transformer is connected to a resistor of resistance R_2. What is the resistance seen from the primary terminals? From Eqs. (4.40) and (4.41),

$$R_1 = \frac{v_1}{i_1} = R_2\left(\frac{N_1}{N_2}\right)^2 \tag{4.42}$$

© 2006 by Taylor & Francis Group, LLC

Of course, if the primary voltage is sinusoidal, complex notation can be used, and the resistances R_1 and R_2 can be replaced by complex impedances Z_1 and Z_2. Finally, if we assume that in an ideal transformer there are no losses, all of the power delivered to the primary can be delivered to a load connected to the secondary. Note that the voltage in both windings is *distributed*, so that there can exist a relatively high voltage between two adjacent layers of turns. This would be irrelevant if, with increasing frequency, this voltage would not result in increasing capacitive currents and deteriorated transformer performance, i.e., basic transformer equations become progressively less accurate with increasing frequency. The frequency at which a transformer becomes useless depends on many factors and cannot be predicted theoretically.

4.4.8. Induced EMF in Loop Antennas

An electromagnetic plane wave is a traveling field consisting of a magnetic and electric field. The magnetic and electric field vectors are mutually perpendicular and perpendicular to the direction of propagation of the wave. The electric field of the wave is, in fact, an induced (only traveling) electric field. Thus, when a small closed wire loop is placed in the field of the wave, there will be an *emf* induced in the loop. *Small* in this context means much smaller than the wave wavelength, and such a loop is referred to as a *loop antenna*. The maximal *emf* is induced if the plane of the loop is perpendicular to the magnetic field of the wave. For a magnetic field of the wave of root-mean-square (rms) value H, a wave frequency f, and loop area (normal to the magnetic field vector) S, the rms value of the *emf* induced in the loop is

$$emf = \left|\frac{d\Phi}{dt}\right| = 2\pi\mu_0 f \cdot H \cdot S \tag{4.43}$$

4.5. EVALUATION OF MUTUAL AND SELF-INDUCTANCE

The simplest method for evaluating mutual and self-inductance is using Eqs. (4.15) and (4.17), provided that the magnetic flux through one of the contours can be calculated. This is possible in some relatively simple, but practical cases. Some of these are presented below.

4.5.1. Examples of Mutual Inductance Calculations

Mutual Inductance Between a Toroidal Coil and a Wire Loop Encircling the Toroid

In order to find the mutual inductance between a contour C_1 and a toroidal coil C_2 with N turns, Fig. 4.16, determining L_{12} is not at all obvious, because the surface of a toroidal coil is complicated. However, $L_{21} = \Phi_{21}/I_2$ is quite simple to find. The flux $d\Phi$ through the surface $dS = h \cdot dr$ in the figure is given by

$$d\Phi_{21}(r) = B(r) \cdot dS = \frac{\mu_0 N I_2}{2\pi r} h \cdot dr \tag{4.44}$$

© 2006 by Taylor & Francis Group, LLC

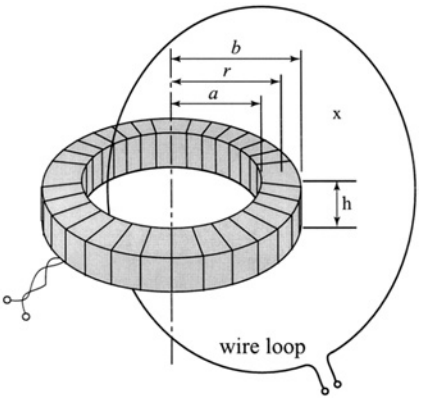


Figure 4.16 A toroidal coil and a single wire loop encircling the toroid.

so that the total flux through C_1, equal to the flux through the cross section of the torus, is

$$\Phi_{21} = \frac{\mu_0 N I_2 h}{2\pi}\int_a^b \frac{dr}{r} = \frac{\mu_0 N I_2 h}{2\pi}\ln\frac{b}{a} \qquad \text{or} \qquad L_{12} = L_{21} = \frac{\mu_0 N h}{2\pi}\ln\frac{b}{a} \tag{4.45}$$

Note that mutual inductance in this case does not depend at all on the shape of the wire loop. Also, if a larger mutual inductance (and thus larger induced *emf*) is required, the loop can simply be wound two or more times around the toroid, to obtain two or more times larger inductance. This is the principle of operation of transformers.

Mutual Inductance Between Two Toroidal Coils

As another example, let us find the mutual inductance between two toroidal coils tightly wound one on top of the other on a core of the form shown in Fig. 4.16. Assume that one coil has N_1 turns and the other N_2 turns. If a current I_2 flows through coil 2, the flux through coil 1 is just N_1 times the flux Φ_{21} from the preceding example, where N should be substituted by N_2. So

$$L_{12} = L_{21} = \frac{\mu_0 N_1 N_2 h}{2\pi}\ln\frac{b}{a} \tag{4.46}$$

Mutual Inductance of Two Thin Coils

Let the mutual inductance of two simple loops be L_{12}. If we replace the two loops by two very thin coils of the same shapes, with N_1 and N_2 turns of very thin wire, the mutual inductance becomes $N_1 N_2 L_{12}$, which is obtained directly from the induced electric field. Similarly, if a thin coil is made of N turns of very thin wire pressed tightly together, its self-inductance is N^2 times that of a single turn of wire.

Mutual Inductance of Two Crossed Two-wire Lines

A two-wire line crosses another two-wire line at a distance d. The two lines are normal. Keeping in mind Eq. (4.3) for the induced electric field, it is easily concluded that their mutual inductance is zero.

© 2006 by Taylor & Francis Group, LLC

4.5.2. Inductors and Examples of Self-inductance Calculations

Self-inductance of a Toroidal Coil

Consider again the toroidal coil in Fig. 4.16. If the coil has N turns, its self-inductance is obtained directly from Eq. (4.45): This flux exists through all the N turns of the coil, so that the flux the coil produces through itself is simply N times that in Eq. (4.45). The self-inductance of the coil in Fig. 4.16 is therefore

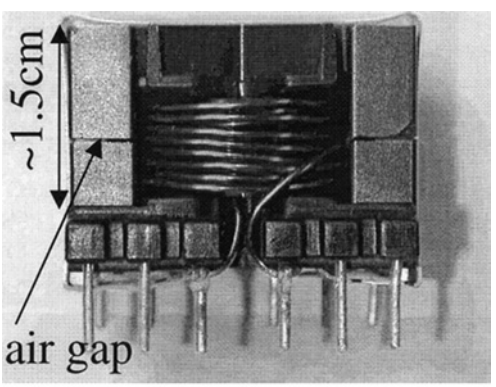
$$L = \frac{\mu_0 N^2 h}{2\pi} \ln \frac{b}{a} \tag{4.47}$$

Self-inductance of a Thin Solenoid

A thin solenoid of length b and cross-sectional area S is situated in air and has N tightly wound turns of thin wire. Neglecting edge effects, the self-inductance of the solenoid is given by

$$L = \frac{\mu_0 N^2 S}{b} \tag{4.48}$$

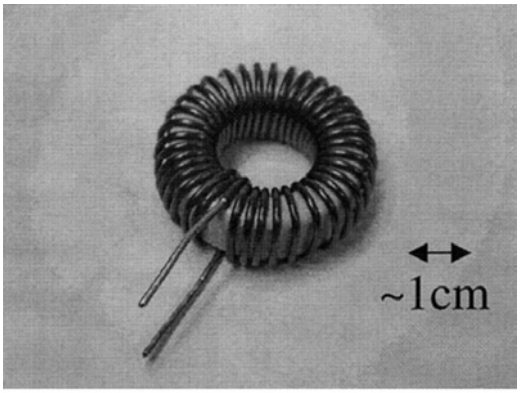
However, in a practical inductor, there exists mutual capacitance between the windings, resulting in a parallel resonant equivalent circuit for the inductor. At low frequencies, the capacitor is an open circuit, but as frequency increases, the reactance of the capacitor starts dominating. At resonance, the parallel resonant circuit is an open, and beyond that frequency, the inductor behaves like a capacitor. To increase the valid operating range for inductors, the windings can be made smaller, but that limits the current handling capability. Figure 4.17 shows some examples of inductor implementations.



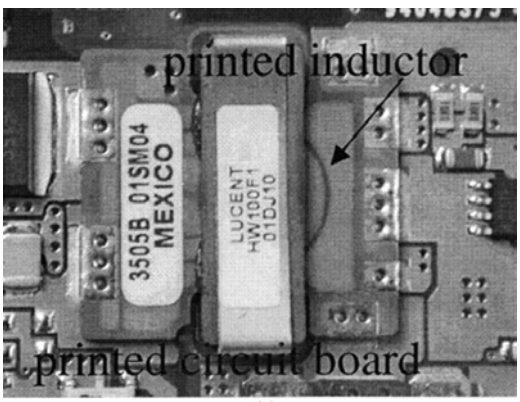
(a)

Figure 4.17 (a) A low-frequency inductor, with $L \approx 1$ mH, wound on a core with an air gap, (c) an inductor with a permalloy core with $L \approx 0.1$ mH, (c) a printed inductor surrounded by a ferromagnetic core with $L \approx 10\,\mu$H, (d) small higher frequency inductors with $L \approx 1\,\mu$H, (e) a chip inductor for surface-mount circuits up to a few hundred MHz with $L \approx 0.1\,\mu$H, and (f) a micromachined spiral inductor with $L \approx 10$–20 nH and a cutoff frequency of 30 GHz ($Q > 50$).

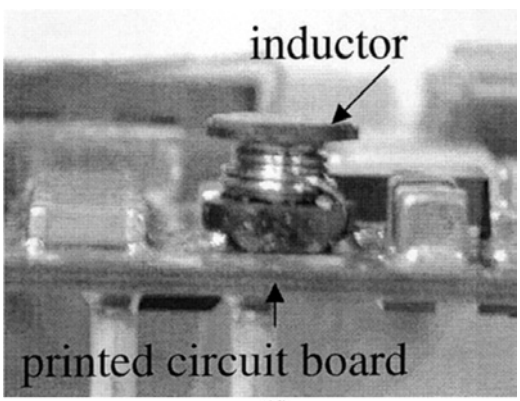
© 2006 by Taylor & Francis Group, LLC



(b)



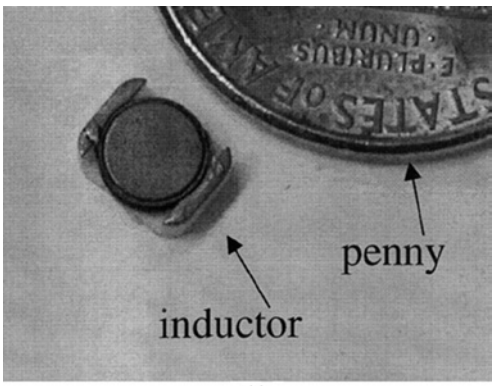
(c)



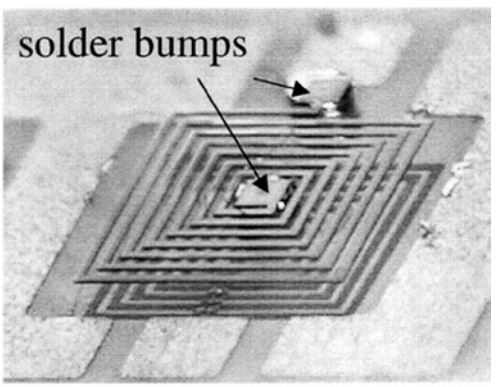
(d)

Figure 4.17 Continued.

© 2006 by Taylor & Francis Group, LLC



(e)



(f)

Figure 4.17 Continued.

The inductor in Fig. 4.17a is wound on a ferromagnetic core, which has the effect of increasing the inductance by μ_r. The small air gap in the core increases the current handling capability of the inductor. To see why this is true, the magnetic circuit equations (3.31)–(3.33) can be applied, using the approximation that an effective μ_r can be defined for the ferromagnetic. Let the effective length (perimeter of the centerline) of the core be L and that of the air gap $L_0 \ll L$. Then

$$NI = R_{m0}\Phi_0 + R_m\Phi \cong (R_{m0} + R_m)\Phi = \frac{1}{\mu_0}B\left(L_0 + \frac{L}{\mu_r}\right) \tag{4.49}$$

where N is the number of windings of the inductor, Φ is the magnetic flux through the core and, approximately, through the air gap, and B is the corresponding magnetic flux density. For an inductor wound on a core without an air gap, the above expression becomes

$$NI' = R_m\Phi = \frac{L}{\mu_0\mu_r}B \tag{4.50}$$

© 2006 by Taylor & Francis Group, LLC

where I' is the current with no air gap in the core, assuming the number of windings, the flux and the dimensions of the core are kept the same. The ratio of the two currents is

$$\frac{I}{I'} = \frac{L_0 + L/\mu_r}{L/\mu_r} = 1 + \frac{\mu_r L_0}{L} \tag{4.51}$$

It can be seen that for the same windings, flux, size, and core material, a higher current in the windings can be used if an air gap is present. Typical relative permeabilities of ferromagnetic cores are in the several thousands (see Table 3.1), while the air gap length is controlled by insulator (usually mylar) sheets of variable thickness, on the order of a fraction of a millimeter. Almost all the magnetic energy is contained in the air gap, since the magnetic field in the gap is μ_r times larger than in the core. The gap therefore enables both larger inductance values and higher current handling, as long as the ferromagnetic does not saturate.

In moly permalloy materials (see Table 3.1), the relative permeability is smaller than in pure ferromagnetics because the material is made with distributed air gaps (bubbles). An inductor with a permalloy core is shown in Fig. 4.17b. The distributed air regions increase the magnetic energy and therefore the current handling capability. Since the effective μ_r is lower, the inductance values are not as high (in the μH range).

Figure 4.17c shows a printed spiral inductor, whose value is increased by wrapping a core around the printed-circuit board. Such inductors can have values on the order of 0.1 mH. Figure 4.15d shows a small high-frequency inductor (several hundred MHz) with a value on the order of tens of microhenry, and Fig. 4.17e shows a surface-mount inductor with values on the order of 0.1 μH and cutoff frequency in the few hundred megahertz range. Figure 4.17f shows a miniature high-frequency micromachined (MEM) inductor suspended in air in order to reduce capacitance due to the presence of the dielectric, resulting in values of inductance on the order of 10–20 nH with a usable frequency range above 20 GHz [21]. At high frequencies, due to the skin effect, the loss in the inductor becomes large, and values of the Q factor are in the range of $Q > 0$.

Self-inductance of a Coaxial Cable

Let us find the external self-inductance per unit length of a coaxial cable. We first need to figure out through which surface to find the flux. If the cable is connected to a generator at one end and to a load at the other, the current flows "in" through the inner conductor and flows back through the outer conductor. The flux through such a contour, for a cable of length h, is the flux through the rectangular surface in Fig. 4.18,

$$\Phi = \int_a^b B(r)h \cdot dr = \frac{\mu_0 I h}{2\pi} \ln \frac{b}{a} \tag{4.52}$$

so that the external self-inductance per unit length of the cable is

$$L' = \frac{\mu_0}{2\pi} \ln \frac{b}{a} \tag{4.53}$$

As a numerical example, for $b/a = e = 2.71828\ldots$, $L' = 0.2\,\mu\text{H/m}$. For a common high-frequency coaxial cable RG-55/U cable, $a = 0.5$ mm, $b = 2.95$ mm, and the inductance per unit length is around 3.55 nH/cm.

© 2006 by Taylor & Francis Group, LLC

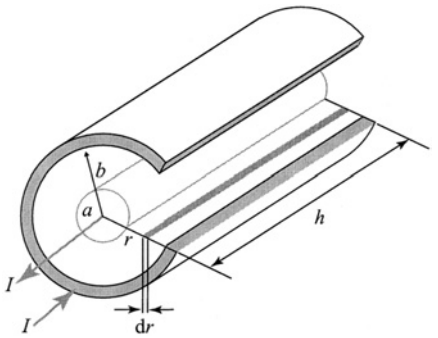


Figure 4.18 Calculating the self-inductance of a coaxial cable.

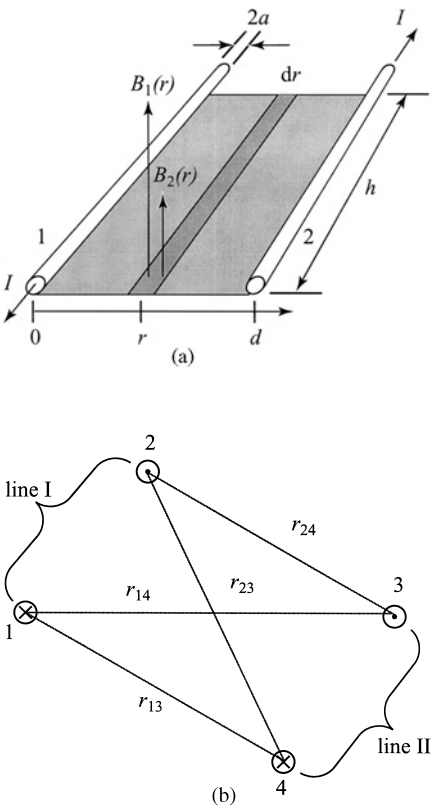


Figure 4.19 (a) Calculating the self-inductance of a thin two-wire line and (b) the mutual inductance between two parallel two-wire lines.

External Self-inductance of a Thin Two-wire Line

A frequently used system for transmission of signals is a thin two-wire line, Fig. 4.19a. Its inductance per unit length is determined as follows. We can imagine that the line is actually a very long rectangular contour (closed with a load at one end, and a generator at the other end), and that we are looking at only one part of it, sketched in the figure.

© 2006 by Taylor & Francis Group, LLC

At a distance r from conductor 1, the current in it produces a magnetic flux density of intensity $B_1(r) = \mu_0 I/2\pi r$, and the current in conductor 2 a magnetic flux density $B_2(r) = \mu_0 I/2\pi(d-r)$. The total flux through a strip of width dr and length h shown in the figure is therefore

$$\Phi = \int_a^{d-a} [B_1(r) + B_2(r)] \cdot h \cdot dr \cong \frac{\mu_0 I h}{\pi} \ln \frac{d}{a} \tag{4.54}$$

since $d \gg a$. The inductance per unit length of the two-wire line is therefore

$$L' = \frac{\mu_0}{\pi} \ln \frac{d}{a} \tag{4.55}$$

As a numerical example, for $d/a = 200$, $L' = 2.12\,\mu\text{H/m}$. We have only calculated the flux through the surface outside of the conductors. The expression for L' above is therefore called the *external self-inductance* of the line. There is also an *internal self-inductance*, due to the flux through the wires themselves (see the example below, Sec. 4.6.3).

Bifilar Coil

To obtain a resistive wire with the smallest self-inductance possible, the wire is bent sharply in the middle and the two mutually insulated halves are pressed tightly together. This results in the smallest external flux possible and, consequently, in the smallest self-inductance. If such a bent wire is wound into a winding, a *bifilar coil* is obtained.

4.5.3. More on Mutual Inductance

Mutual Inductance Between Two Parallel Two-wire Lines

Mutual inductance per unit length of two two-wire lines running parallel to each other, shown in the cross section in Fig. 4.19b, can be obtained by calculating the magnetic flux per unit length due to current in one line through the other. For the reference directions of the two lines the indicated result is

$$L'_{\text{I, II}} = \frac{\Phi'_{\text{I, II}}}{I_{\text{I}}} = \frac{\mu_0}{2\pi} \ln \frac{r_{14} r_{23}}{r_{13} r_{24}} \tag{4.56}$$

Self-inductance and Mutual Inductance of Two Windings over a Toroidal Core

A thin toroidal core of permeability μ, mean radius R, and cross-sectional area S is densely wound with two coils of thin wire, with N_1 and N_2 turns, respectively. The windings are wound one over the other. The self- and mutual inductances of the coils are

$$L_1 = \frac{\mu N_1^2 S}{2\pi R}, \qquad L_2 = \frac{\mu N_2^2 S}{2\pi R}, \qquad L_{12} = \frac{\mu N_1 N_2 S}{2\pi R} \tag{4.57}$$

so that the coupling coefficient $k = 1$.

© 2006 by Taylor & Francis Group, LLC

4.5.4. Neumann's Formula for Inductance Calculations

Neumann's Formula for Mutual Inductance of Two-wire Loops

Starting from the induced electric field due to a thin-wire loop, it is possible to derive a general formula for two thin-wire loops in a homogeneous medium, Fig. 4.20, known as *Neumann's formula*. With reference to Fig. 4.20, it is of the form

$$L_{12} = \frac{\mu_0}{4\pi} \oint_{C_2} \oint_{C_1} \frac{d\mathbf{l}_1 \cdot d\mathbf{l}_2}{r} \tag{4.58}$$

Note that L_{21} would have the same form, except that the order of integration and the dot product of current elements would exchange places. Since this does not affect the result, we conclude that $L_{12} = L_{21}$. Note also that explicit evaluation of the dual integral in Eq. (4.58) can be performed only in rare instances, but it can always be integrated numerically with ease.

Flat Multiconductor Cable

As an example of application of Neumann's formula, consider n narrow coplanar strips that run parallel to each another over a distance d, Fig. 4.21. This is a model of a flat

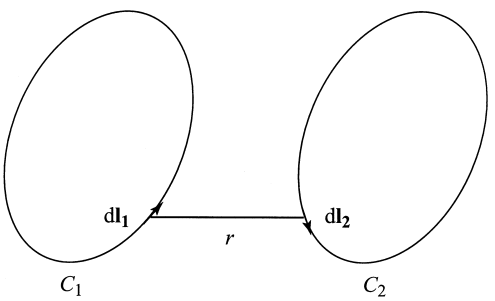


Figure 4.20 Two loops made out of thin wire.

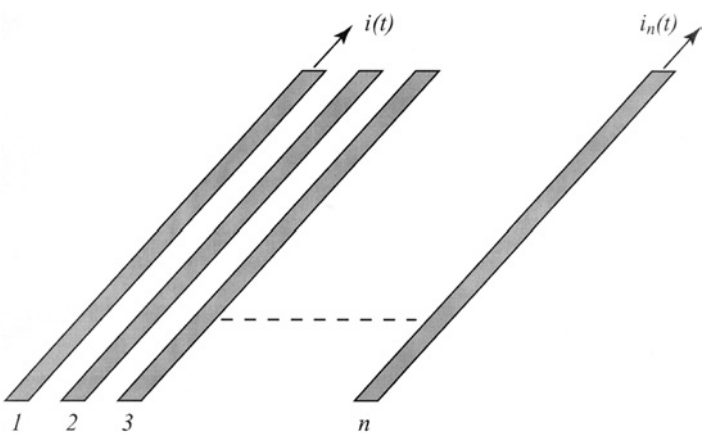


Figure 4.21 A flat multiconductor cable (transmission line).

© 2006 by Taylor & Francis Group, LLC

multiconductor cable (transmission line), such as the ones used to connect a printer to a computer. Let the currents in the strips be $i_1(t), i_2(t), \ldots, i_n(t)$. We wish to determine the *emf* induced, for example, in strip no. 1 by the time-varying current in all the other strips. Although, as explained, this type of coupling is usually referred to as *magnetic coupling*, it is actually an example of mutual coupling by means of an induced electric field.

Note that we need not have closed loops in the Neumann formula—what matters is the induced electric field and the length of the wire in which we determine the induced *emf* (i.e., the line integral of the induced field). The total *emf* in conductor no. 1 induced by the other conductors running parallel to it for a distance d is given by

$$e_1(t) = -\frac{\mu_0}{4\pi}\sum_{j=2}^{n}\frac{di_j(t)}{dt}\left(\int_0^d\int_0^d \frac{dl_1 dl_j}{r_{1j}}\right) \tag{4.59}$$

The elements $dl_1, dl_2, \ldots, dl_n$ are along the center lines of the strips. The integrals can be evaluated explicitly using tables of integrals. Note that the reference direction of currents in all the strips is assumed to be the same.

Neumann's Formula for External Self-inductance of a Wire Loop

Neumann's formula in Eq. (4.58) can be modified to enable the evaluation of external self-inductance. At first glance, one can just consider the case when the two contours, C_1 and C_2, in Neumann's formula overlap, and the self-inductance of a loop results. This is not so, however, because the integral becomes singular and divergent ($1/r$ is zero when elements dl_1 and dl_2 coincide).

To alleviate this problem, assume instead that one loop, e.g., $C = C_1$, is along the axis of the loop and the other, $C' = C_2$, is along the surface of the wire (Fig. 4.22). The distance between line elements of such two contours is never zero, and the integral becomes convergent. The flux computed in this case is the flux that a line current along the loop axis produces through a contour on the wire surface, so this is precisely the external loop inductance. The Neumann formula for the external inductance of a loop is thus

$$L = \frac{\mu_0}{4\pi}\oint_C\oint_{C'}\frac{d\mathbf{l}\cdot d\mathbf{l}'}{r} \tag{4.60}$$

As with Eq. (4.58), it is possible to integrate the dual integral in this equation explicitly only rarely, but it can always be integrated numerically.

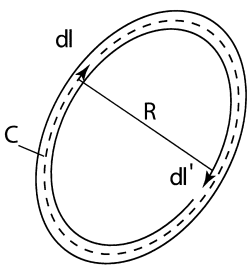


Figure 4.22 A wire loop and two possible contours of integration in the Neumann formula for self-inductance.

© 2006 by Taylor & Francis Group, LLC

4.6. ENERGY AND FORCES IN THE MAGNETIC FIELD: IMPLICATIONS AND APPLICATIONS

4.6.1. Magnetic Energy of Two Magnetically Coupled Contours

In the case of two contours ($n = 2$), Eqs. (4.21) and (4.22) for the magnetic energy of n contours become

$$W_m = \frac{1}{2}(I_1\Phi_1 + I_2\Phi_2) \tag{4.61}$$

and

$$W_m = \frac{1}{2}L_{11}I_1^2 + \frac{1}{2}L_{22}I_2^2 + L_{12}I_1I_2 \tag{4.62}$$

This energy can be smaller or larger than the sum of energies of the two contours when isolated, since L_{12} can be positive or negative.

4.6.2. Losses in Ferromagnetic Materials Due to Hysteresis and Eddy Currents

Let us observe what happens to energy needed to maintain a sinusoidal magnetic field in a piece of ferromagnetic material. The hysteresis curve of the material is shown in Fig. 4.23, and the arrows show the direction in which the operating point is moving in the course of time. According to Eq. (4.24), the energy density that needs to be spent at a point where the magnetic field is H, in order to change the magnetic flux density by dB, is equal to $H\,dB$. In the diagram in Fig. 4.23, this is proportional to the area of the small shaded rectangle. So, the integral of $H\,dB$ is proportional to the sum of all such rectangles as the point moves around the hysteresis curve.

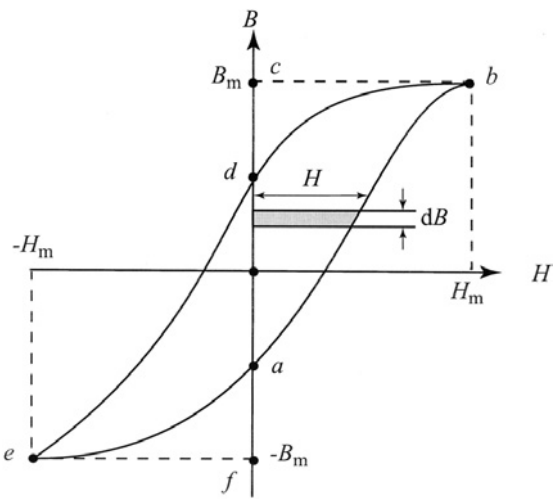


Figure 4.23 Hysteresis curve of a ferromagnetic material.

© 2006 by Taylor & Francis Group, LLC

Starting from point a in Fig. 4.23 moving to point b, the magnetic field H is positive. The increase dB is also positive, so $H.dB$ is positive, and the energy density needed to move from point a to b is proportional to the area of the curved triangle abc in the figure. From b to d, H is positive, but B is decreasing, so that dB is negative. Therefore, the product $H.dB$ is negative, which means that in this region the energy *used up* on maintaining the field is negative. This in turn means that this portion of the energy is returned back from the field to the sources. The density of this returned energy is proportional to the area of the curved triangle bdc. From d to e, the product $H.dB$ is positive, so this energy is spent on maintaining the field, and from e to a, the product is negative, so this energy is returned to the sources. Therefore, we come to the conclusion that only the energy density proportional to the area of the curved triangles bcd and efa is returned to the sources. All the rest, which is proportional to the area formed by the hysteresis loop, is lost to heat in the ferromagnetic material. These losses are known as *hysteresis losses*. If the frequency of the field is f, the operating point circumscribes the loop f times per second. Consequently, *hysteresis losses are proportional to frequency* (and to the volume of the ferromagnetic material if the field is uniform).

If the ferromagnetic material is conductive, there are also eddy-current losses, proportional to the square of frequency. As an example, consider a solenoid with a ferromagnetic core made of thin, mutually insulated sheets. To estimate the eddy-current and hysteresis losses, the *total* power losses are measured at two frequencies, f_1 and f_2, for the same amplitude of the magnetic flux density. The total power losses are found to be P_1 and P_2, respectively. Knowing that hysteresis losses are proportional to frequency, and eddy-current losses to the square of frequency, it is possible to separately determine these losses as follows. First the total losses can be expressed as

$$P = P_{\text{total losses}} = P_{\text{hysterisis losses}} + P_{\text{eddy-current losses}} = Af + Bf^2 \tag{4.63}$$

where A and B are constants. Consequently,

$$P_1 = Af_1 + Bf_1^2 \qquad \text{and} \qquad P_2 = Af_2 + Bf_2^2 \tag{4.64}$$

from which

$$A = \frac{P_1 f_2^2 - P_2 f_1^2}{f_1 f_2 (f_2 - f_1)} \tag{4.65}$$

Ferrite Anechoic Chambers for EMC/EMI Testing

For testing electromagnetic compatibility and interference over a broad frequency range, dimensions of absorber material for adequately low reflections would be very large and impractical. Anechoic chambers made for this purpose have walls made of ferrite material, with very high magnetic losses in the megahertz and gigahertz frequency range. As a result of the high losses, the walls can be made much thinner, and for increased absorption over a broader bandwidth the ferrite tiles are sometimes backed with a dielectric layer of dielectric backed with metal.

© 2006 by Taylor & Francis Group, LLC

4.6.3. Internal Inductance of a Straight Wire at Low Frequencies

The energy of a wire with a current i is distributed outside the wire, as well as inside the wire, since there is a magnetic field both outside and inside the wire. From the energy expression $W_m = Li^2/2$ for a single current contour, we can write

$$L_{\text{internal}} = \frac{2W_{m\ \text{inside conductor}}}{i^2} \qquad \text{and} \qquad L_{\text{external}} = \frac{2W_{m\ \text{outside conductor}}}{i^2} \tag{4.66}$$

Consider a long straight wire of circular cross section and permeability μ. If the current in the wire is assumed to be distributed uniformly (or very nearly so, i.e., we consider low frequencies), according to Ampère's law, the magnetic field inside the wire is equal to $H(r) = Ir/2\pi a^2$. Using Eqs. (4.26) and (4.66), we find that the internal inductance of the wire per unit length is given by

$$L'_{\text{internal}} = \frac{\mu_0}{8\pi} \tag{4.67}$$

Note that the internal inductance does not depend on the radius of the wire.

4.6.4. Total Inductance of a Thin Two-wire Line at Low Frequencies

The total self-inductance per unit length of a thin two-wire line with wires made of a material with permeability μ, radius of the wires a, and distance between the wire axes $d \gg a$ is the sum of its external inductance in Eq. (4.55) and the internal *inductance of both wires*:

$$L' = L'_{\text{internal}} + L'_{\text{external}} = \frac{\mu_0}{\pi}\ln\frac{d}{a} + 2\frac{\mu_0}{8\pi} \tag{4.68}$$

As a numerical example, if $\mu = \mu_0$ and $d/a = 100$, we get $L'_{\text{external}} = 1.84\,\mu\text{H/m}$ and $L'_{\text{internal}} = 0.1\,\mu\text{H/m}$. In this example, the external inductance is much larger than the internal inductance. This is usually the case.

4.6.5. Force of an Electromagnet

As an example of the force formula in Eq. (4.27), the attractive force of an electromagnet, sketched in Fig. 4.24, is evaluated below. The electromagnet is in the shape of a horseshoe, and its magnetic force is lifting a weight W, shown in the figure. This is a magnetic circuit. Let us assume that when the weight W moves by a small amount dx upward, the flux in the magnetic circuit does not change. That means that when the weight is moved upward, the only change in magnetic energy is the *reduction* in energy contained in the two air gaps, due to their decreased length. This energy reduction is

$$-\frac{dW_m}{dx} = \frac{1}{2}\frac{B^2}{\mu_0}2S \cdot dx \tag{4.69}$$

and the force is now equal to

$$F_x = \frac{1}{2}\frac{B^2}{\mu_0}2S = \frac{\Phi^2}{\mu_0}S \tag{4.70}$$

© 2006 by Taylor & Francis Group, LLC

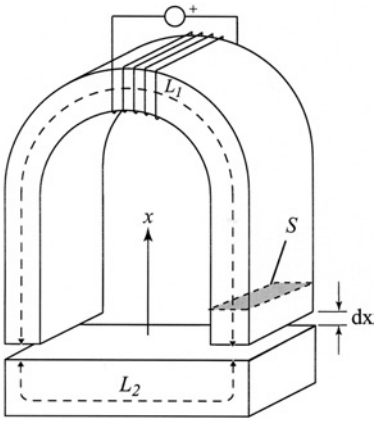


Figure 4.24 Sketch of an electromagnet lifting a ferromagnetic weight.

As a numerical example, let $B = 1\,\mathrm{T}$, and $S = 1000\,\mathrm{cm}^2$. For this case, $F_x = 7.96 \times 10^4\,\mathrm{N}$, which means that this electromagnet can lift a weight of about 8 tons! Such electromagnets are used, for example, in cranes for lifting large pieces of iron.

4.6.6. Comparison of Electric and Magnetic Pressure

The expression for the pressure of magnetic forces on boundary surfaces between materials of different magnetic properties can be obtained starting from Eqs. (4.27) and (4.28). For two magnetic media of permeabilities μ_1 and μ_2, the pressure on the interface, assumed to be directed into medium 1, is given by

$$p = \frac{1}{2}(\mu_2 - \mu_1)\left(H_{\tan g}^2 + \frac{B_{\mathrm{norm}}^2}{\mu_1 \mu_2}\right) \tag{4.71}$$

with the reference direction of pressure into medium 1. We know that magnetic flux density of about 1 T is quite large and not easily attainable. Therefore, for $B_{\mathrm{norm}} = 1\,\mathrm{T}$, $H_{\tan g} = 0$, and $\mu_2 \gg \mu_1 = \mu_0$, the largest magnetic pressure that can be obtained is on the order of

$$p_{m,\max} \approx 400{,}000\ \frac{\mathrm{N}}{\mathrm{m}^2}$$

The electric pressure on a metallic conductor in vacuum is given by the expression $p_{e,\max} = (1/2)\varepsilon_0 E^2$. The electric strength of air is about $3 \times 10^6\,\mathrm{V/m}$. This means that the largest electric pressure in air is approximately

$$p_{e,\max} = 0.5(8.86 \times 10^{-12})\,(3 \times 10^6)^2 \simeq 40\,\mathrm{N/m}^2$$

Consequently, the ratio of the maximal magnetic and maximal electric pressure is approximately

$$\frac{p_{m,\max}}{p_{e,\max}} = 10{,}000$$

© 2006 by Taylor & Francis Group, LLC

This is an interesting and important conclusion. Although "electric" and "magnetic" versions of almost any device can be designed using electric and magnetic forces, the magnetic version will require much smaller space for the same amount of power. To get an idea for the order of magnitude of the magnetic and electric pressure, note that a typical car-tire pressure is around $200\,\text{kPa} = 200{,}000\,\text{N/m}^2$ (or 30 psi).

4.6.7. High-frequency Resistance and Internal Inductance of a Wire

It is easy to understand from Ohm's law that a metal wire has a resistance at dc given by the resistance for a uniform resistor. As the frequency increases, this resistance changes due to the skin effect, i.e., the redistribution of current across the cross section of the conductor. For a cylindrical wire of radius a, the associated resistance per unit length is given by $R' = R_S/2\pi a$, where $R_S = \sqrt{\omega\mu/2\sigma}$ is the surface resistance of the conductor with conductivity σ at an angular frequency ω and is obtained from assuming the current flows through a cross section determined by skin depth.

At high frequencies, a wire also has magnetic energy stored nonuniformly inside it, and this is associated with internal inductance of the wire per unit length. It can be shows that the reactive power at high frequencies inside a conductor is equal to the power of Joule losses due to the wire surface resistance. The power of heat loss is given by $P'_{\text{heat}} = R_S I^2/2\pi a$, and the inductance per unit length of a cylindrical wire at high frequencies is found from $R_S I^2/2\pi a = \omega L'_{\text{int}} I^2$. Therefore, the resistance and internal inductance per unit length of a cylindrical metal wire at a frequency ω are given by

$$R' = \frac{1}{2\pi a}\sqrt{\frac{\omega\mu}{2\sigma}} \qquad \text{and} \qquad L'_{\text{int}} = \frac{1}{2\pi a\omega}\sqrt{\frac{\omega\mu}{2\sigma}} \tag{4.72}$$

This frequency-variable internal inductance should be added to the external inductance when, e.g., calculating the characteristic impedance of cables at high frequencies.

4.7. SOME INTERESTING EXAMPLES OF ELECTROMAGNETIC INDUCTION

In this section a few interesting examples that the authors encountered in practice, and that they feel might be useful to the reader, are described. In particular, a commonly encountered case is that of signal cross-talk due to a current-carrying wire that passes through a hole in a metal casing (Sec. 4.7.2).

4.7.1. Mutual Inductance Between Monophase Cables Laid on the Bottom of the Sea

Assume we have three single-phase 60-Hz power cables laid at the bottom of the sea, for example, to supply electric power to an island. The cables are spaced by a few hundred meters and are parallel to each other. (Three distant single-phase instead of one three-phase cable are often used for safety reasons: if a ship accidentally pulls and breaks one cable with an anchor, two are left. In addition, usually a spare single-phase cable is laid to enable quick replacement of a damaged one.) If the length of the cables is long (in practice, it can be many kilometers), we might expect very large mutual inductance between these

© 2006 by Taylor & Francis Group, LLC

cables, due to the huge loops they form, and, consequently, unbalanced currents in the three cables. The 60-Hz sea water skin depth, however, tells us that there will be practically no mutual inductance between the cables.

4.7.2. Cross Talk Due to Current in Wire Passing Through a Hole in a Metal Casing

An interesting and commonly encountered practical effect occurs when a single wire with a high-frequency current passes through a hole in a metal sheet (e.g., the side of a metal chassis). *High frequency* in this case means that skin depth should be much smaller than the sheet thickness. Consequently, the reasoning is valid, surprisingly, also for power frequencies (60 or 50 Hz) if the sheet is ferromagnetic and its thickness is on the order of 10 mm or greater.

If Ampère's law is applied to a contour encircling the hole so that it is further away from the sheet surface than the skin depth, the line integral of vector **H** is practically zero, because there is no magnetic field so deep in the sheet. This means that the total current encircled by the contour is practically zero, i.e., *that the current induced on the hole surface is practically the same as the current in the wire*. Of course, once it leaves the surface of the hole, this current continues to flow over both sheet surfaces, producing its own magnetic and induced electric field. Consequently, the signal carried by the current through the wire can be transmitted as described to undesirable places, causing unexpected cross talk.

4.7.3. Rough Calculation of Induced Voltages in a Human Body Due to Currents in Power Lines

There is often concern that fields radiated by power lines might be harmful for human health. It is interesting to do a calculation of the induced voltages in the human body that result from currents in power lines. There are two mechanisms by which a voltage can be produced in such a situation: that produced by the electric field, and that induced through electromagnetic induction due to magnetic field variations. This example shows a calculation of the two mechanisms on the example of the human head, assuming that it is the most important, and possibly the most sensitive part of a human body. Assume that a human head is a sphere with a radius of 10 cm and consisting mostly of salty water. For this example, the induced voltages in the head are calculated for the power lines being as close as 20 m from the human head, and they carry 100 A of unbalanced current, Fig. 4.25. For any other input information, the results can be easily scaled.

The magnetic flux density 20 m away from a wire with 100 A of current is $B = \mu_0 I/2\pi r = 1\,\mu\text{T}$. (How large is this? The earth's dc magnetic field is on average 50 μT on the surface, and as a person moves in this field, some voltage will be induced, but humans are presumably adapted to this effect.) Faraday's law can be used to calculate the induced *emf* around the head due to the calculated value of B:

$$\oint_{\text{around head}} \mathbf{E} \cdot d\mathbf{l} = -\frac{\partial}{\partial t} \int_{\text{head cross section}} \mathbf{B} \cdot d\mathbf{S}$$

In complex notation, the above equation becomes $2\pi E_{\text{induced}} = -j\omega B \pi a^2$, where a is the radius of the head. From here, the value of the voltage due to the induced field

© 2006 by Taylor & Francis Group, LLC

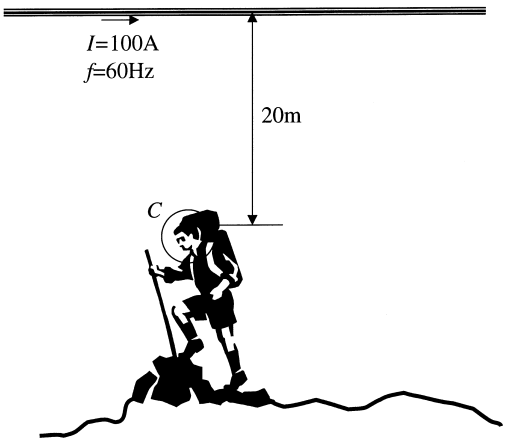


Figure 4.25 Calculating approximately the electromagnetic influence of a power line on the head of a human standing under it.

across a single 10-μm cell in our head is calculated to be about 33 pV for a power-line frequency of 60 Hz.

This is only one component of the effect of power lines on the human. The other is due to the electric field, which depends on the voltage of the power line. A reasonable value for the electric field close to the power line is around $E = 1\,\text{kV/m}$. Salt water has a resistivity ρ of about 1 Ωm, and to find the voltage across a single cell that can be added to the induced voltage, the following reasoning can be made. We first find the charge density σ produced on the head due to the high field, we then find the total charge Q by integration. Assuming a 60-Hz field frequency, this changing charge will produce a current I, and a corresponding current density J. The current density in the nonperfect conductor produces an ohmic voltage drop across a cell. The following equations describe this reasoning, assuming the head is perfectly spherical:

$$\sigma(\theta) = 3\varepsilon_0 E\cos\theta \qquad Q = \int_{\text{head}} \sigma\, dS = \int_0^{\pi/2} \sigma(\theta) 2\pi a \sin\theta \cdot d\theta = 3\pi\varepsilon_0 E a^2$$

$$I = 2\pi f Q = 0.315\,\mu\text{A} \qquad J = \frac{I}{\pi a^2} \qquad V = E \cdot 10\,\mu\text{m} = \rho J \cdot 10\,\mu\text{m} \simeq 100\,\text{pV}$$

Thus the total voltage across a cell in the human head due to a high-voltage line nearby is calculated to be about 133 pV. For comparison, normal neural impulses are much larger: they are spikes with around 100-mV amplitudes, frequency between 1 and 100 Hz, and duration of about a millisecond.

REFERENCES

1. Beaty, H.B.; Kirtley, J.L. Jr. *Electric Motor Handbook*; McGraw-Hill: New York, 1998.
2. Becker, R. *Electromagnetic Fields and Interactions*; Dover Publications: New York, 1982.
3. Bewley, L.V. *Flux Linkages and Electromagnetic Induction*; Dover Publications: New York, 1964.

© 2006 by Taylor & Francis Group, LLC

4. Cheston, W.B. *Elementary Theory of Electromagnetic Fields*; John Wiley & Sons: New York, 1964.
5. Coren, R.L. *Basic Engineering Electromagnetics: An Applied Aprroach*; Prentice Hall: Upper Saddle River, NJ, 1989.
6. Coulson, C.A. *Electricity*, Oliver and Boyd, Edingurgh, 1953.
7. Elliot, R.S. *Electromagnetics*; McGraw-Hill: New York, 1996.
8. Harsanyi, G. *Sensors in Biomedical Applications: Fundamentals, Technology and Applications*; Technomic Pub. Co.: Lancaster, Pa, 2000.
9. Hayt, W.H. Jr. *Engineering Electromagnetics*; McGraw-Hill: New York, 1967.
10. Iskander, M.F. *Electromagnetic Fields and Waves*; Prentice Hall: Upper Saddle River, NJ, 1992.
11. Jordan, E.C.; Balmain, K.G. *Electromagnetic Waves and Radiating Systems*; Prentice Hall: Upper Saddle River, NJ, 1968.
12. Küpfmüler, K. *Einführung in die theoretische Elektrotechnik*; Springer Verlag: Berlin, 1962.
13. Landau, L.; Lifschitz, L. *The Classical Theory of Fields*; Addison-Wesley: Reading, MA, 1951.
14. Marshall, S.V.; DuBroff, R.E.; Skitek G.G. *Electromagnetic Concepts and Applications*; 4th Ed.; Prentice Hall: Upper Saddle River, NJ, 1996.
15. Maxwell, J.C. *A Treatise on Electricity and Magnetism*; Dover Publications: New York, 1954; Vols. 1 and 2.
16. Popović, B.D. *Introductory Engineering Electromagnetics*; Addison-Wesley: Reading, MA, 1971.
17. Popović, R.S. *Hall Effect Devices*; The Adam Hilger Series on Sensors; IOP Publishing, 1991.
18. Popović, Z.; Popović, B.D. *Introductory Electromagnetics*; Prentice Hall: Upper Saddle River, NJ, 1999.
19. Ramo, S.; Whinnery, J.R.; van Duzer, T. *Fields and Waves in Communication Electronics*; John Wiley & Sons: New York, 1965; 1st Ed.; and 1994; 3rd Ed.
20. Ramshaw, R.; van Heeswijk, R.G. *Energy Conversion: Electric Motors and Generators*; Saunders College Publishing: Philadelphia, 1990.
21. Ripka, P., Ed. *Magnetic Sensors and Magnetometers*; Boston: Artech House, 2001.
22. Rutledge, D.B. *Electromagnetics*; Lecture Notes, Caltech, 1990.
23. Schelkunoff, S.A. *Electromagnetic Fields*; Blaisdell Publishing Company: New York, 1963.
24. Smythe, W.R. *Static and Dynamic Electricity*; McGraw-Hill: New York, 1968.
25. Sommerfeld, A. *Electrodynamics*; Academic Press: New York, 1952.
26. "Electric Motors", Encyclopedia Britannica, 2004. Encyclopedia Britannica Online, 16 Feb. 2004. <http://www.search.eb.com/eb/article?eu=108542>

© 2006 by Taylor & Francis Group, LLC

5
Wave Propagation

Mohammad Kolbehdari
Intel Corporation
Hillsboro, Oregon

Matthew N. O. Sadiku
Prairie View A&M University
Prairie View, Texas

Electromagnetic (EM) wave propagation deals with the transfer of energy or information from one point (a transmitter) to another (a receiver) through the media such as material space, transmission line, and waveguide. It can be described using both theoretical models and practical models based on empirical results. Here we describe the free-space propagation model, path loss models, and the empirical path loss formula. Before presenting these models, we first discuss the theoretical basis and characteristics of EM waves as they propagate through material media.

5.1. WAVE EQUATIONS AND CHARACTERISTICS

The EM wave propagation theory can be described by Maxwell's equations [1,2].

$$\nabla \times \mathbf{E}(\mathbf{r},t) = -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r},t) \tag{5.1}$$

$$\nabla \times \mathbf{H}(\mathbf{r},t) = \frac{\partial}{\partial t}\mathbf{D}(\mathbf{r},t) + \mathbf{J}(\mathbf{r},t) \tag{5.2}$$

$$\nabla \cdot \mathbf{D}(\mathbf{r},t) = \rho(\mathbf{r},t) \tag{5.3}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r},t) = 0 \tag{5.4}$$

In the above equations, the field quantities **E** and **H** represent, respectively, the electric and magnetic fields, and **D** and **B** the electric and magnetic displacements. **J** and ρ represent the current and charge sources. This set of differential equations relates the time and space rates of change of various field quantities at a point in space and time. Furthermore, the position vector **r** defines a particular location in space (x,y,z) at which the field is being measured. Thus, for example,

$$\mathbf{E}(x,y,z,t) = \mathbf{E}(\mathbf{r},t) \tag{5.5}$$

© 2006 by Taylor & Francis Group, LLC

An auxiliary relationship between the current and charge densities, $\mathbf{J}$ and ρ, called the *continuity equation* is given by

$$\nabla \cdot \mathbf{J}(\mathbf{r},t) = -\frac{\partial}{\partial t}\rho(\mathbf{r},t) \tag{5.6}$$

The constitutive relationships between the field quantities and electric and magnetic displacements provide the additional constraints needed to solve Eqs. (5.1) and (5.2). These equations characterize a given isotropic material on a macroscopic level in terms of two scalar quantities as

$$\mathbf{B} = \mu\mathbf{H} = \mu_0\mu_r\mathbf{H} \tag{5.7}$$

$$\mathbf{D} = \varepsilon\mathbf{E} = \varepsilon_0\varepsilon_r\mathbf{E} \tag{5.8}$$

where $\mu_0 = 4\pi \times 10^{-7}$ H/m (henrys per meter) is the permeability of free space and $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m (farads per meter) is the permittivity of free space. Also, ε_r and μ_r, respectively, characterize the effects of the atomic and molecular dipoles in the material and the magnetic dipole moments of the atoms constituting the medium.

Maxwell's equations, given by Eqs. (5.1) to (5.4), can be simplified if one assumes time-harmonic fields, i.e., fields varying with a sinusoidal frequency ω. For such fields, it is convenient to use the complex exponential $e^{j\omega t}$. Applying the time-harmonic assumption to Eqs. (5.1) to (5.4), we obtain the time-harmonic wave propagation equations

$$\nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mathbf{B}(\mathbf{r}) \tag{5.9}$$

$$\nabla \times \mathbf{H}(\mathbf{r}) = j\omega\mathbf{D}(\mathbf{r}) + \mathbf{J}(\mathbf{r}) \tag{5.10}$$

$$\nabla \cdot \mathbf{D}(\mathbf{r}) = \rho(\mathbf{r}) \tag{5.11}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}) = 0 \tag{5.12}$$

The solution of Maxwell's equations in a source free isotropic medium can be obtained by using Eqs. (5.9) and (5.10) and applying Eqs. (5.7) and (5.8) as follows:

$$\nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mu\mathbf{H}(\mathbf{r}) \tag{5.13}$$

$$\nabla \times \mathbf{H}(\mathbf{r}) = j\omega\varepsilon\mathbf{E}(\mathbf{r}) \tag{5.14}$$

Taking the curl of the Eq. (5.13) and using Eq. (5.14) we get

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mu\nabla \times \mathbf{H}(\mathbf{r}) = \omega^2\mu\varepsilon\mathbf{E}(\mathbf{r}) \tag{5.15}$$

Using a vector identity, and noting that $\rho = 0$, we can write Eq. (15) as

$$\nabla^2\mathbf{E}(\mathbf{r}) + \omega^2\mu\varepsilon\mathbf{E}(\mathbf{r}) = 0 \tag{5.16}$$

This relation is called the *wave equation*. For example, the x component of $\mathbf{E}(\mathbf{r})$ is

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)E_x(\mathbf{r}) + \omega^2\mu\varepsilon E_x(\mathbf{r}) = 0 \tag{5.17}$$

© 2006 by Taylor & Francis Group, LLC

5.1.1. Attenuation

If we consider the general case of a lossy medium that is charge free ($\rho = 0$), Eqs. (5.9) to (5.12) can be manipulated to yield Helmholz' wave equations

$$\nabla^2 \mathbf{E} - \gamma^2 \mathbf{E} = 0 \tag{5.18}$$

$$\nabla^2 \mathbf{H} - \gamma^2 \mathbf{H} = 0 \tag{5.19}$$

where $\gamma = \alpha + j\beta$ is the propagation constant, α is the attenuation constant in nepers per meter or decibels per meter, and β is the phase constant in radians per meters. Constants α and β are given by

$$\alpha = \omega \sqrt{\frac{\mu\varepsilon}{2}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\varepsilon}\right)^2} - 1\right]} \tag{5.20}$$

$$\beta = \omega \sqrt{\frac{\mu\varepsilon}{2}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\varepsilon}\right)^2} + 1\right]} \tag{5.21}$$

where $\omega = 2\pi f$ is the angular frequency of the wave and σ is the conductivity of the medium.

Without loss of generality, if we assume that the wave propagates in the z direction and the wave is polarized in the x direction, solving the wave equations in Eqs. (5.18) and (5.19), we obtain

$$E_x = E_0 e^{-\alpha z} \cos(\omega t - \beta z) \tag{5.22}$$

$$H_y = \frac{E_0}{|\eta|} e^{-\alpha z} \cos(\omega t - \beta z - \theta_\eta) \tag{5.23}$$

where $\eta = |\eta| \angle \theta_\eta$ is the intrinsic impedance of the medium and is given by

$$|\eta| = \frac{\sqrt{\mu/\varepsilon}}{\sqrt[4]{\left[1 + (\sigma/\omega\varepsilon)^2\right]}} \qquad \tan 2\theta_\eta = \frac{\sigma}{\omega\varepsilon} \qquad 0 \le \theta_\eta \le 45^\circ \tag{5.24}$$

Equations (5.22) and (5.23) show that as the EM wave propagates in the medium, its amplitude is attenuated to $e^{-\alpha z}$.

5.1.2. Dispersion

A plane electromagnetic wave can be described as

$$E_x(\mathbf{r}) = E_{x0} e^{-j\mathbf{k}\cdot\mathbf{r}} = E_{x0} e^{-j(k_x x + k_y y + k_z z)} \tag{5.25}$$

where E_{x0} is an arbitrary constant, and $\mathbf{k} = k_x \mathbf{a}_x + k_y \mathbf{a}_y + k_z \mathbf{a}_z$ is the vector wave and $\mathbf{r} = \mathbf{a}_x x + \mathbf{a}_y y + \mathbf{a}_z z$ is the vector observation point. The substitution of the assumed form

© 2006 by Taylor & Francis Group, LLC

of the plane wave in Eq. (5.17) yields

$$k_x^2 + k_y^2 + k_z^2 = k^2 = \omega^2 \mu \varepsilon \tag{5.26}$$

This equation is called the *dispersion relation*. It may also be written in terms of the velocity v defined by

$$k = \frac{\omega}{v} \tag{5.27}$$

The other components of $\mathbf{E}(\mathbf{r})$ with the same wave equation also have the same dispersion equation.

The characteristic impedance of plane wave in free space is given by

$$\eta = \frac{|\mathbf{E}|}{|\mathbf{H}|} = \sqrt{\frac{\mu}{\varepsilon}} = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 377\,\Omega \tag{5.28}$$

5.1.3. Phase Velocity

By assuming $k = k_z = \omega\sqrt{\mu\varepsilon}$, the electric field can be described by

$$E(z,t) = E_0 \cos(\omega t - k_z z + \varphi) \tag{5.29}$$

For an observer moving along with the same velocity as the wave, an arbitrary point on the wave will appear to be constant, which requires that the argument of the $E(z,t)$ be constant as defined by

$$\omega t - k_z z + \varphi = \text{constant} \tag{5.30}$$

Taking the derivative with respect to the z yields

$$\frac{dz}{dt} = \frac{\omega}{k_z} = v_p \tag{5.31}$$

where v_p is defined as the phase velocity; for free space it is

$$\frac{\omega}{k_z} = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} \cong 3 \times 10^8 \text{ m/s} \tag{5.32}$$

which is the velocity of light in free space.

5.1.4. Group Velocity

A signal consisting of two equal-amplitude tones at frequencies $\omega_0 \pm \Delta\omega$ can be represented by

$$f(t) = 2\cos\omega_0 t \cos\Delta\omega t \tag{5.33}$$

© 2006 by Taylor & Francis Group, LLC

which corresponds to a signal carrier at frequency ω_0 being modulated by a slowly varying envelope having the frequency $\Delta\omega$. If we assume that each of the two signals travels along a propagation direction z with an associated propagation constant $k(\omega)$, then the propagation constant of each signal is $k(\omega_0 \pm \Delta\omega)$. An expansion in a first-order Taylor series yields

$$k(\omega_0 \pm \Delta\omega) \cong k(\omega_0) \pm \Delta\omega k^1(\omega_0) \tag{5.34}$$

where

$$k^1(\omega_0) = \frac{dk(\omega)}{d\omega}|_{\omega=\omega_0} \tag{5.35}$$

The substitution of Eq. (5.34) into Eq. (5.33) following some mathematical manipulation yields

$$f(t,z) = 2\cos\omega_0(t-\tau_p)\ \cos\Delta\omega(t-\tau_g) \tag{5.36}$$

where

$$\tau_p = \frac{k(\omega_0)}{\omega_0} z \tag{5.37}$$

and

$$\tau_g = k^1(\omega_0)z \tag{5.38}$$

The quantities τ_p and τ_g are defined as the phase and group delays, respectively. The corresponding propagation velocities are

$$v_p = \frac{z}{\tau_p} \tag{5.39}$$

$$v_g = \frac{z}{\tau_g} \tag{5.40}$$

For a plane wave propagating in a uniform unbounded medium, the propagation constant is a linear function of frequency given in Eq. (5.26). Thus, for a plane wave, phase and group velocities are equal and given by

$$v_p = v_g = \frac{1}{\sqrt{\mu\varepsilon}} \tag{5.41}$$

It is worthwhile to mention that if the transmission medium is a waveguide, $k(\omega)$ is no longer a linear function of frequency. It is very useful to use the ω-k diagram shown in Fig. 5.1, which plots ω versus $k(\omega)$. In this graph, the slope of a line drawn from the origin to the frequency ω_0 gives the phase velocity and the slope of the tangent to the curve at ω_0 yields the group velocity.

© 2006 by Taylor & Francis Group, LLC

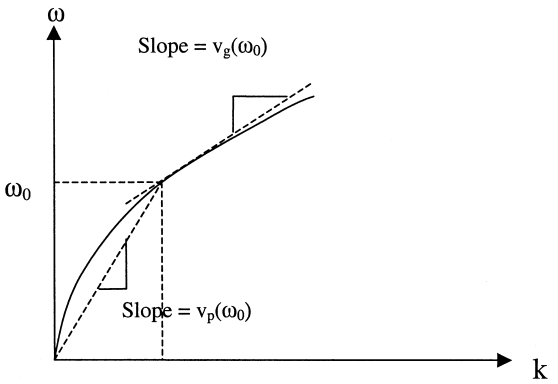


Figure 5.1 ω-k diagram.

5.1.5. Polarization

The electric field of a plane wave propagating in the z direction with no components in the direction of propagation can be written as

$$\mathbf{E}(z) = (\mathbf{a}_x E_{x0} + \mathbf{a}_y E_{y0}) e^{-jk_z z} \tag{5.42}$$

By defining

$$E_{x0} = E_{x0} e^{j\varphi_x} \tag{5.43}$$

$$E_{y0} = E_{y0} e^{j\varphi_y} \tag{5.44}$$

we obtain

$$E(z) = (\mathbf{a}_x E_{x0} e^{j\varphi_x} + \mathbf{a}_y E_{y0}^{j\varphi_y}) e^{-jk_z z} \tag{5.45}$$

Assuming $A = E_{y0}/E_{x0}$ and $\varphi = \varphi_y - \varphi_x$, and $E_{x0} = 1$, we can write Eq. (5.45) as

$$\mathbf{E}(z) = (\mathbf{a}_x + \mathbf{a}_y A e^{j\varphi}) e^{-jk_z z} \tag{5.46}$$

Case I: $A = 0$. $\mathbf{E}(z) = \mathbf{a}_x e^{-jk_z z}$ and $\mathbf{E}(z,t) = \mathbf{a}_x \cos(\omega t - k_z z)$. The movement of the electric field vector in the $z = 0$ plane is along the x axis. This is known as a *linearly polarized wave* along the x axis.

Case II: $A = 1$, $\varphi = 0$.

$$\mathbf{E}(z) = (\mathbf{a}_x + \mathbf{a}_y) e^{-jk_z z} \tag{5.47}$$

and

$$\mathbf{E}(z,t) = (\mathbf{a}_x + \mathbf{a}_y) \cos(\omega t - k_z z) \tag{5.48}$$

© 2006 by Taylor & Francis Group, LLC

This is again a linear polarized wave with the electric field vector at 45 degrees with respect to the x axis.

Case III: $A = 2$, $\varphi = 0$.

$$\mathbf{E}(z,t) = (\mathbf{a}_x + 2\mathbf{a}_y)\ \cos(\omega t - k_z z) \tag{5.49}$$

This is again a linear polarized wave with the electric field vector at 63 degrees with respect to the x axis.

Case IV: $A = 1$, $\varphi = \pi/2$.

$$\mathbf{E}(z) = (\mathbf{a}_x + j\mathbf{a}_y)e^{-jk_z z} \tag{5.50}$$

and

$$\mathbf{E}(z,t) = \mathbf{a}_x \cos(\omega t - k_z z) - \mathbf{a}_y \sin(\omega t - k_z z) \tag{5.51}$$

In this case the electric field vector traces a circle and the wave is defined to be left-handed circularly polarized. Similarly, with $\varphi = -\pi/2$, it is a right-handed circularly polarized wave.

Case VI: $A = 2$ and $\varphi \neq 0$. This is an example of an elliptically polarized wave.

5.1.6. Poynting's Theorem

The relationships between the electromagnetic fields can be described by Poynting's theorem. For an isotropic medium, Maxwell's curl equations can be written as

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \tag{5.52}$$

$$\nabla \times \mathbf{H} = \varepsilon \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J} \tag{5.53}$$

where the current density $\mathbf{J}$ can be described as having two components:

$$\mathbf{J} = \mathbf{J}_s + \mathbf{J}_c \tag{5.54}$$

where $\mathbf{J}_c = \sigma\mathbf{E}$ represents conduction current density induced by the presence of the electric fields and $\mathbf{J}_s$ is a source current density that induces electromagnetic fields. The quantity $\mathbf{E} \cdot \mathbf{J}$ has the unit of power per unit volume (watts per unit cubic meter). From Eqs. (5.52) and (5.53) we can get

$$\mathbf{E} \cdot \mathbf{J} = \mathbf{E} \cdot \nabla \times \mathbf{H} - \varepsilon \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} \tag{5.55}$$

Applying the vector identity

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot \nabla \times \mathbf{A} - \mathbf{A} \cdot \nabla \times \mathbf{B} \tag{5.56}$$

© 2006 by Taylor & Francis Group, LLC

gives

$$\mathbf{E} \cdot \mathbf{J} = \mathbf{H} \cdot \nabla \times \mathbf{E} - \nabla \cdot (\mathbf{E} \times \mathbf{H}) - \varepsilon \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} \tag{5.57}$$

Substituting Eq. (5.52) into Eq. (5.57) yields

$$\mathbf{E} \cdot \mathbf{J} = -\mu \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} - \nabla \cdot (\mathbf{E} \times \mathbf{H}) - \varepsilon \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} \tag{5.58}$$

Integrating Eq. (5.58) over an arbitrary volume V that is bounded by surface S with an outward unit normal to the surface $\hat{\mathbf{n}}$ shown in Fig. 5.2 gives

$$\iiint_v \mathbf{E} \cdot \mathbf{J}\, dv = \frac{\partial}{\partial t}\left(\iiint_v 1/2\mu |\mathbf{H}|^2\, dv + \iiint_v 1/2\varepsilon |\mathbf{E}|^2\, dv \right) + \oiint_s \hat{\mathbf{n}} \cdot (\mathbf{E} \times \mathbf{H})\, ds \tag{5.59}$$

where the following identity has been used

$$\iiint_v \nabla \cdot \mathbf{A}\, dv = \oiint_s \hat{\mathbf{n}} \cdot \mathbf{A}\, ds \tag{5.60}$$

Equation (5.59) represents the Poynting theorem. The terms $1/2\mu|\mathbf{H}|^2$ and $1/2\varepsilon|\mathbf{E}|^2$ are the energy densities stored in magnetic and electric fields, respectively. The term $\oiint_s \hat{\mathbf{n}} \cdot (\mathbf{E} \times \mathbf{H})\, ds$ describes the power flowing out of the volume V. The quantity $\mathbf{P} = \mathbf{E} \times \mathbf{H}$ is called the *Poynting vector* with the unit of power per unit area. For example, the Poynting theorem can be applied to the plane electromagnetic wave given in Eq. (5.29), where $\varphi = 0$. The wave equations are

$$E_x(z,t) = E_0 \cos(\omega t - k_z z) \tag{5.61}$$

$$H_y(z,t) = \sqrt{\frac{\varepsilon}{\mu}} E_0 \cos(\omega t - k_z z) \tag{5.62}$$

The Poynting vector is in the z direction and is given by

$$P_z = E_x H_y = \sqrt{\frac{\varepsilon}{\mu}} E_0^2 \cos^2(\omega t - k_z z) \tag{5.63}$$

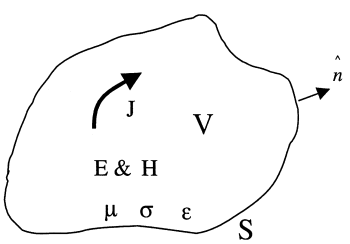


Figure 5.2 A volume V enclosed by surface S and unit vector $\mathbf{n}$.

© 2006 by Taylor & Francis Group, LLC

Applying the trigonometric identity yields

$$P_z = \sqrt{\frac{\varepsilon}{\mu}} E_0^2 \left[\frac{1}{2} + \frac{1}{2} \cos 2\,(\omega t - k_z z) \right] \tag{5.64}$$

It is worth noting that the constant term shows that the wave carries a time-averaged power density and there is a time-varying portion representing the stored energy in space as the maxima and the minima of the fields pass through the region.

We apply the time-harmonic representation of the field components in terms of complex phasors and use the time average of the product of two time-harmonic quantities given by

$$\langle A(t)B(t) \rangle = \tfrac{1}{2} \mathrm{Re}(AB^*) \tag{5.65}$$

where B^* is the complex conjugate of B. The time average Poynting power density is

$$\langle \mathbf{P} \rangle = \tfrac{1}{2} \mathrm{Re}(\mathbf{E} \times \mathbf{H}^*) \tag{5.66}$$

where the quantity $\mathbf{P} = \mathbf{E} \times \mathbf{H}$ is defined as the complex Poynting vector.

5.1.7. Boundary Conditions

The boundary conditions between two materials shown in Fig. 5.3 are

$$E_{t1} = E_{t2} \tag{5.67}$$

$$H_{t1} = H_{t2} \tag{5.68}$$

In the vector form, these boundary conditions can be written as

$$\hat{\mathbf{n}} \times (\mathbf{E}_1 - \mathbf{E}_2) = 0 \tag{5.69}$$

$$\hat{\mathbf{n}} \times (\mathbf{H}_1 - \mathbf{H}_2) = 0 \tag{5.70}$$

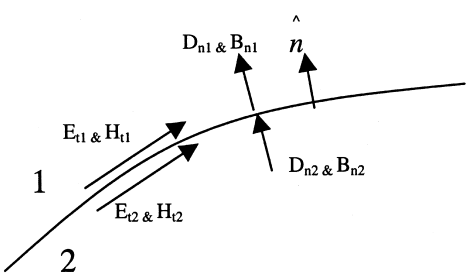


Figure 5.3 Boundary conditions between two materials.

© 2006 by Taylor & Francis Group, LLC

Thus, the tangential components of electric and magnetic field must be equal on the two sides of any boundary between the physical media. Also for a charge- and current-free boundary, the normal components of electric and magnetic flux density are continuous, i.e.,

$$D_{n1} = D_{n2} \tag{5.71}$$

$$B_{n1} = B_{n2} \tag{5.72}$$

For the perfect conductor (infinite conductivity), all the fields inside of the conductor are zero. Thus, the continuity of the tangential electric fields at the boundary yields

$$E_t = 0 \tag{5.73}$$

Since the magnetic fields are zero inside of the conductor, the continuity of the normal magnetic flux density yields

$$B_n = 0 \tag{5.74}$$

Furthermore, the normal electric flux density is

$$D_n = \rho_s \tag{5.75}$$

where ρ_s is a surface charge density on the boundary. The tangential magnetic field is discontinuous by the current enclosed by the path, i.e.,

$$H_t = \mathbf{J}_s \tag{5.76}$$

where $\mathbf{J}_s$ is the surface current density.

5.1.8. Wave Reflection

We now consider the problem of a plane wave obliquely incident on a plane interface between two lossless dielectric media, as shown in Fig. 5.4. It is conventional to define two cases of the problem: the electric field is in the xz plane (parallel polarization) or normal to the xz plane (parallel polarization). Any arbitrary incident plane wave may be treated as a linear combination of the two cases. The two cases are solved in the same manner: obtaining expressions for the incident, reflection, and transmitted fields in each region and matching the boundary conditions to find the unknown amplitude coefficients and angles.

For parallel polarization, the electric field lies in the xz plane so that the incident fields can be written as

$$E_i = E_0(a_x \cos\theta_i - a_z \sin\theta_i)e^{-jk_1(x\sin\theta_i + z\cos\theta_i)} \tag{5.77}$$

$$H_i = \frac{E_0}{\eta_1} a_y e^{-jk_1(x\sin\theta_i + z\cos\theta_i)} \tag{5.78}$$

© 2006 by Taylor & Francis Group, LLC

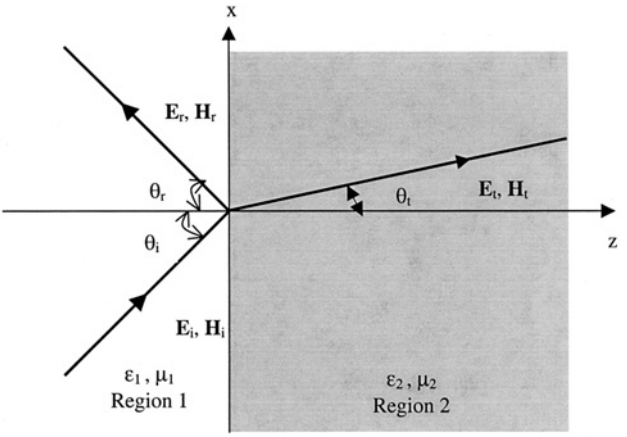


Figure 5.4 A plane wave obliquely incident at the interface between two regions.

where $k_1 = \omega\sqrt{\mu_1\varepsilon_1}$ and $\eta_1 = \sqrt{\mu_1/\varepsilon_1}$. The reflected and transmitted fields can be obtained by imposing the boundary conditions at the interface.

$$E_r = \Gamma_{||} E_0 (a_x \cos\theta_r + a_z \sin\theta_r) e^{-jk_1(x\sin\theta_r - z\cos\theta_r)} \tag{5.79}$$

$$H_r = -\frac{\Gamma_{||} E_0}{\eta_1} a_y e^{-jk_1(x\sin\theta_r - z\cos\theta_r)} \tag{5.80}$$

$$E_t = E_0 T_{||} (a_x \cos\theta_t - a_z \sin\theta_t) e^{-jk_2(x\sin\theta_t + z\cos\theta_t)} \tag{5.81}$$

$$H_t = \frac{E_0 T_{||}}{\eta_2} a_y e^{-jk_2(x\sin\theta_t + z\cos\theta_t)} \tag{5.82}$$

where $k_2 = \omega\sqrt{\mu_2\varepsilon_2}$, $\eta_2 = \sqrt{\mu_2/\varepsilon_2}$

$$\theta_r = \theta_i \qquad k_1 \sin\theta_i = k_2 \sin\theta_t \quad \text{(Snell's law)} \tag{5.83}$$

$$\Gamma_{||} = \frac{\eta_2 \cos\theta_t - \eta_1 \cos\theta_i}{\eta_2 \cos\theta_t + \eta_1 \cos\theta_i} \tag{5.84}$$

and

$$T_{||} = \frac{2\eta_2 \cos\theta_i}{\eta_2 \cos\theta_t + \eta_1 \cos\theta_i} \tag{5.85}$$

For perpendicular polarization, the electric field is normal to the xz plane. The incident fields are given by

$$E_i = E_0 a_y e^{-jk_1(x\sin\theta_i + z\cos\theta_i)} \tag{5.86}$$

$$H_i = \frac{E_0}{\eta_1}(-a_x \cos\theta_i + a_z \sin\theta_i) e^{-jk_1(x\sin\theta_i + z\cos\theta_i)} \tag{5.87}$$

© 2006 by Taylor & Francis Group, LLC

while the reflected and transmitted fields are

$$E_r = \Gamma_\perp E_0 a_y e^{-jk_1(x\sin\theta_r - z\cos\theta_r)} \tag{5.88}$$

$$H_r = \frac{\Gamma_\perp E_0}{\eta_1}(a_x\cos\theta_r + a_z\sin\theta_r)e^{-jk_1(x\sin\theta_r - z\cos\theta_r)} \tag{5.89}$$

$$E_t = E_0 T_\perp a_y e^{-jk_2(x\sin\theta_t + z\cos\theta_t)} \tag{5.90}$$

$$H_t = \frac{E_0 T_\perp}{\eta_2}(-a_x\cos\theta_t + a_z\sin\theta_t)e^{-jk_2(x\sin\theta_t + z\cos\theta_t)} \tag{5.91}$$

where

$$k_1\sin\theta_i = k_1\sin\theta_r = k_2\sin\theta_t \qquad \text{(Snell's law)} \tag{5.92}$$

$$\Gamma_\perp = \frac{\eta_2\cos\theta_i - \eta_1\cos\theta_t}{\eta_2\cos\theta_i + \eta_1\cos\theta_t} \tag{5.93}$$

and

$$T_\perp = \frac{2\eta_2\cos\theta_i}{\eta_2\cos\theta_i + \eta_1\cos\theta_t} \tag{5.94}$$

5.2. FREE-SPACE PROPAGATION MODEL

The free-space propagation model is used in predicting the received signal strength when the transmitter and receiver have a clear line-of-sight path between them. If the receiving antenna is separated from the transmitting antenna in free space by a distance r, as shown in Fig. 5.5, the power received P_r by the receiving antenna is given by the Friis equation [3]

$$P_r = G_r G_t\left(\frac{\lambda}{4\pi r}\right)^2 P_t \tag{5.95}$$

where P_t is the transmitted power, G_r is the receiving antenna gain, G_t is the transmitting antenna gain, and λ is the wavelength ($= c/f$) of the transmitted signal. The Friis equation



Figure 5.5 Basic wireless system.

© 2006 by Taylor & Francis Group, LLC

relates the power received by one antenna to the power transmitted by the other, provided that the two antennas are separated by $r > 2d^2/\lambda$, where d is the largest dimension of either antenna. Thus, the Friis equation applies only when the two antennas are in the far field of each other. It also shows that the received power falls off as the square of the separation distance r. The power decay as $1/r^2$ in a wireless system, as exhibited in Eq. (5.95), is better than the exponential decay in power in a wired link. In actual practice, the value of the received power given in Eq. (5.95) should be taken as the maximum possible because some factors can serve to reduce the received power in a real wireless system. This will be discussed fully in the next section.

From Eq. (5.95), we notice that the received power depends on the product P_tG_t. The product is defined as the *effective isotropic radiated power* (EIRP), i.e.,

$$\text{EIRP} = P_tG_t \tag{5.96}$$

The EIRP represents the maximum radiated power available from a transmitter in the direction of maximum antenna gain relative to an isotropic antenna.

5.3. PATH LOSS MODEL

Wave propagation seldom occurs under the idealized conditions assumed in Sec. 5.1. For most communication links, the analysis in Sec. 5.1 must be modified to account for the presence of the earth, the ionosphere, and atmospheric precipitates such as fog, raindrops, snow, and hail [4]. This will be done in this section.

The major regions of the earth's atmosphere that are of importance in radio wave propagation are the troposphere and the ionosphere. At radar frequencies (approximately 100 MHz to 300 GHz), the troposphere is by far the most important. It is the lower atmosphere consisting of a nonionized region extending from the earth's surface up to about 15 km. The ionosphere is the earth's upper atmosphere in the altitude region from 50 km to one earth radius (6370 km). Sufficient ionization exists in this region to influence wave propagation.

Wave propagation over the surface of the earth may assume one of the following three principal modes:

Surface wave propagation along the surface of the earth
Space wave propagation through the lower atmosphere
Sky wave propagation by reflection from the upper atmosphere

These modes are portrayed in Fig. 5.6. The sky wave is directed toward the ionosphere, which bends the propagation path back toward the earth under certain conditions in a limited frequency range (below 50 MHz approximately). This is highly dependent on the condition of the ionosphere (its level of ionization) and the signal frequency. The surface (or ground) wave takes effect at the low-frequency end of the spectrum (2–5 MHz approximately) and is directed along the surface over which the wave is propagated. Since the propagation of the ground wave depends on the conductivity of the earth's surface, the wave is attenuated more than if it were propagation through free space. The space wave consists of the direct wave and the reflected wave. The direct wave travels from the transmitter to the receiver in nearly a straight path while the reflected wave is due to ground reflection. The space wave obeys the optical laws in that direct and reflected wave

© 2006 by Taylor & Francis Group, LLC

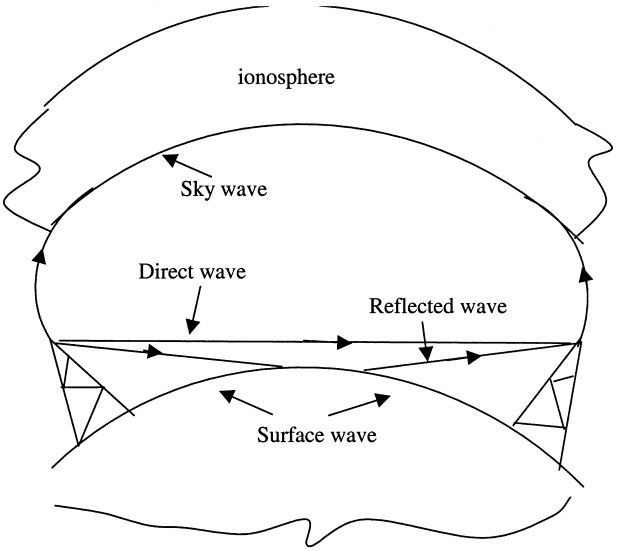


Figure 5.6 Modes of wave propagation.

components contribute to the total wave component. Although the sky and surface waves are important in many applications, we will only consider space wave in this chapter.

In case the propagation path is not in free space, a correction factor F is included in the Friis equation, Eq. (5.74), to account for the effect of the medium. This factor, known as the *propagation factor*, is simply the ratio of the electric field intensity E_m in the medium to the electric field intensity E_o in free space, i.e.,

$$F = \frac{E_m}{E_o} \tag{5.97}$$

The magnitude of F is always less than unity since E_m is always less than E_o. Thus, for a lossy medium, Eq. (5.95) becomes

$$P_r = G_r G_t \left(\frac{\lambda}{4\pi r}\right)^2 P_t |F|^2 \tag{5.98}$$

For practical reasons, Eqs. (5.95) and (5.98) are commonly expressed in logarithmic form. If all the terms are expressed in decibels (dB), Eq. (5.98) can be written in the logarithmic form as

$$P_r = P_t + G_r + G_t - L_o - L_m \tag{5.99}$$

where P = power in dB referred to 1 W (or simply dBW), G = gain in dB, L_o = free-space loss in dB, and L_m loss in dB due to the medium. (Note that $G\,\text{dB} = 10\log_{10} G$.) The free-space loss is obtained directly from Eq. (5.98) as

$$L_o = 20\log\frac{4\pi r}{\lambda} \tag{5.100}$$

© 2006 by Taylor & Francis Group, LLC

while the loss due to the medium is given by

$$L_m = -20\log|F| \tag{5.101}$$

Our major concern in the rest of this subsection is to determine L_o and L_m for an important case of space propagation that differs considerably from the free-space conditions.

The phenomenon of multipath propagation causes significant departures from free-space conditions. The term *multipath* denotes the possibility of EM wave propagating along various paths from the transmitter to the receiver. In multipath propagation of an EM wave over the earth's surface, two such path exists: a direct path and a path via reflection and diffractions from the interface between the atmosphere and the earth. A simplified geometry of the multipath situation is shown in Fig. 5.7. The reflected and diffracted component is commonly separated into two parts: one *specular* (or coherent) and the other *diffuse* (or incoherent), that can be separately analyzed. The specular component is well defined in terms of its amplitude, phase, and incident direction. Its main characteristic is its conformance to Snell's law for reflection, which requires that the angles of incidence and reflection be equal and coplanar. It is a plane wave, and as such, is uniquely specified by its direction. The diffuse component, however, arises out of the random nature of the scattering surface and, as such, is nondeterministic. It is not a plane wave and does not obey Snell's law for reflection. It does not come from a given direction but from a continuum.

The loss factor F that accounts for the departures from free-space conditions is given by

$$F = 1 + \Gamma\rho_s DS(\theta)e^{-j\Delta} \tag{5.102}$$

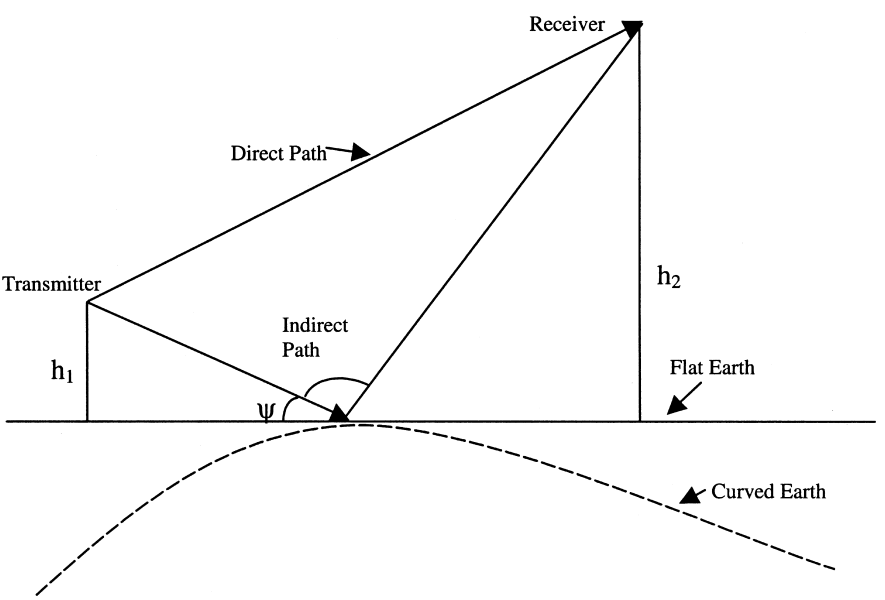


Figure 5.7 Multipath geometry.

© 2006 by Taylor & Francis Group, LLC

where

Γ = Fresnel reflection coefficient.
ρ_s = roughness coefficient.
D = divergence factor.
$S(\theta)$ = shadowing function.
Δ = phase angle corresponding to the path difference.

The Fresnel reflection coefficient Γ accounts for the electrical properties of the earth's surface. Since the earth is a lossy medium, the value of the reflection coefficient depends on the complex relative permittivity ε_c of the surface, the grazing angle ψ, and the wave polarization. It is given by

$$\Gamma = \frac{\sin\psi - z}{\sin\psi + z} \tag{5.103}$$

where

$$z = \sqrt{\varepsilon_c - \cos^2\psi} \qquad \text{for horizontal polarization} \tag{5.104}$$

$$z = \frac{\sqrt{\varepsilon_c - \cos^2\psi}}{\varepsilon_c} \qquad \text{for vertical polarization} \tag{5.105}$$

$$\varepsilon_c = \varepsilon_r - j\frac{\sigma}{\omega\varepsilon_o} = \varepsilon_r - j60\sigma\lambda \tag{5.106}$$

ε_r and σ are, respectively, the dielectric constant and the conductivity of the surface; ω and λ are, respectively, the frequency and wavelength of the incident wave; and ψ is the grazing angle. It is apparent that $0 < |\Gamma| < 1$.

To account for the spreading (or divergence) of the reflected rays due to earth curvature, we introduce the divergence factor D. The curvature has a tendency to spread out the reflected energy more than a corresponding flat surface. The divergence factor is defined as the ratio of the reflected field from curved surface to the reflected field from flat surface. Using the geometry of Fig. 5.8, we get D as

$$D = \left(1 + \frac{2G_1G_2}{a_eG\sin\psi}\right)^{-1/2} \tag{5.107}$$

where $G = G_1 + G_2$ is the total ground range and $a_e = 6370$ km is the effective earth radius. Given the transmitter height h_1, the receiver height h_2, and the total ground range G, we can determine G_1, G_2, and ψ. If we define

$$p = \frac{2}{\sqrt{3}}\left[a_e(h_1 + h_2) + \frac{G^2}{4}\right]^{1/2} \tag{5.108}$$

$$\alpha = \cos^{-1}\left[\frac{2a_e(h_1 - h_2)G}{p^3}\right] \tag{5.109}$$

© 2006 by Taylor & Francis Group, LLC

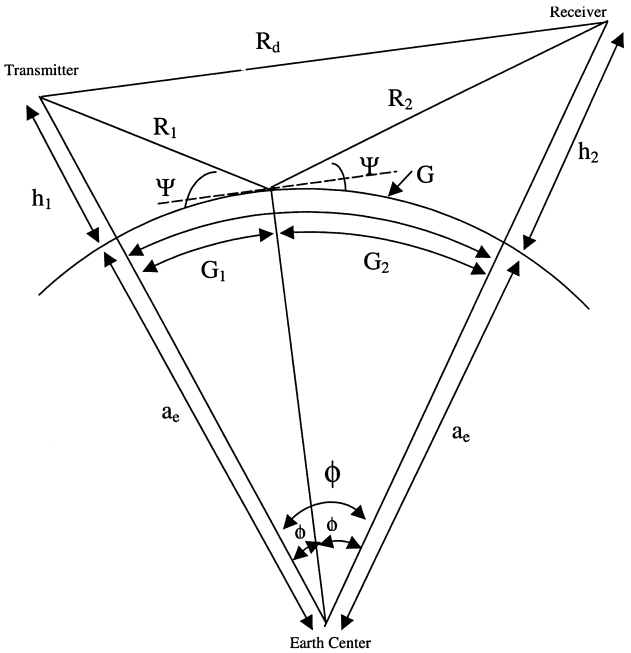


Figure 5.8 Geometry of spherical earth reflection.

and assume $h_1 \leq h_2$ and $G_1 \leq G_2$, using small angle approximation yields [5]

$$G_1 = \frac{G}{2} + p\cos\frac{\pi+\alpha}{3} \tag{5.110}$$

$$G_2 = G - G_1 \tag{5.111}$$

$$\phi_i = \frac{G_i}{a_e}, \qquad i = 1,2 \tag{5.112}$$

$$R_i = \left[h_i^2 + 4a_e(a_e + h_i)\sin^2(\phi_i/2)\right]^{1/2} \qquad i = 1,2 \tag{5.113}$$

The grazing angle is given by

$$\psi = \sin^{-1}\left(\frac{2a_e h_1 + h_1^2 - R_1^2}{2a_e R_1}\right) \tag{5.114}$$

or

$$\psi = \sin^{-1}\left(\frac{2a_e h_1 + h_1^2 + R_1^2}{2(a_e + h_1)R_1}\right) - \phi_1 \tag{5.115}$$

Although D varies from 0 to 1, in practice D is a significant factor at low grazing angle ψ (less than 0.1 %).

© 2006 by Taylor & Francis Group, LLC

The phase angle corresponding to the path difference between direct and reflected waves is given by

$$\Delta = \frac{2\pi}{\lambda}(R_1 + R_2 - R_d) \tag{5.116}$$

The roughness coefficient ρ_s takes care of the fact that the earth surface is not sufficiently smooth to produce specular (mirrorlike) reflection except at very low grazing angle. The earth's surface has a height distribution that is random in nature. The randomness arises out of the hills, structures, vegetation, and ocean waves. It is found that the distribution of the heights of the earth's surface is usually the gaussian or normal distribution of probability theory. If σ_h is the standard deviation of the normal distribution of heights, we define the roughness parameters

$$g = \frac{\sigma_h \sin\psi}{\lambda} \tag{5.117}$$

If $g < 1/8$, specular reflection is dominant; if $g > 1/8$, diffuse scattering results. This criterion, known as the *Rayleigh criterion*, should only be used as a guideline since the dividing line between a specular and a diffuse reflection or between a smooth and a rough surface is not well defined [6]. The roughness is taken into account by the roughness coefficient $(0 < \rho_s < 1)$, which is the ratio of the field strength after reflection with roughness taken into account to that which would be received if the surface were smooth. The roughness coefficient is given by

$$\rho_s = \exp[-2(2\pi g)^2] \tag{5.118}$$

Shadowing is the blocking of the direct wave due to obstacles. The shadowing function $S(\theta)$ is important at low grazing angle. It considers the effect of geometric shadowing—the fact that the incident wave cannot illuminate parts of the earth's surface shadowed by higher parts. In a geometric approach, where diffraction and multiple scattering effects are neglected, the reflecting surface will consist of well-defined zones of illumination and shadow. As there will be no field on a shadowed portion of the surface, the analysis should include only the illuminated portions of the surface. A pictorial representation of rough surfaces illuminated at angle of incidence $\theta(= 90^\circ - \psi)$ is shown in Fig. 5.9. It is evident

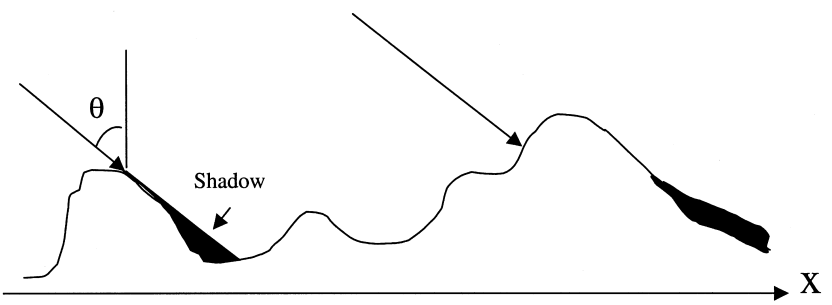


Figure 5.9 Rough surface illuminated at an angle of incidence.

© 2006 by Taylor & Francis Group, LLC

from the figure that the shadowing function $S(\theta)$ equals unity when $\theta = 0$ and zero when $\theta = \pi/2$. According to Smith [7],

$$S(\theta) = \frac{1 - (1/2)\text{erfc}(a)}{1 + 2B} \tag{5.119}$$

where $\text{erfc}(x)$ is the complementary error function,

$$\text{erfc}(x) = 1 - \text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \tag{5.120}$$

and

$$B = \frac{1}{4a}\left[\frac{1}{\sqrt{\pi}} e^{a^2} - a\,\text{erfc}(a)\right] \tag{5.121}$$

$$a = \frac{\cot\theta}{2s} \tag{5.122}$$

$$s = \frac{\sigma_h}{\sigma_l} = \text{rms surface slope} \tag{5.123}$$

In Eq. (5.123), σ_h is the rms roughness height and σ_l is the correlation length. Alternative models for $S(\theta)$ are available in the literature. Using Eqs. (5.103) to (5.123), we can calculate the loss factor in Eq. (5.102). Thus

$$L_o = 20\log\frac{4\pi R_d}{\lambda} \tag{5.124}$$

$$L_m = -20\log\left[1 + \Gamma\rho_s D S(\theta) e^{-j\Delta}\right] \tag{5.125}$$

5.4. EMPIRICAL PATH LOSS FORMULA

Both theoretical and experimental propagation models are used in predicting the path loss. In addition to the theoretical model presented in the previous section, there are empirical models for finding path loss. Of the several models in the literature, the Okumura et al. model [8] is the most popular choice for analyzing mobile-radio propagation because of its simplicity and accuracy. The model is based on extensive measurements in and around Tokyo, compiled into charts, that can be applied to VHF and UHF mobile-radio propagation. The medium path loss (in dB) is given by [9]

$$L_p = \begin{cases} A + B\log_{10} r & \text{for urban area} \\ A + B\log_{10} r - C & \text{for suburban area} \\ A + B\log_{10} r - D & \text{for open area} \end{cases} \tag{5.126}$$

© 2006 by Taylor & Francis Group, LLC

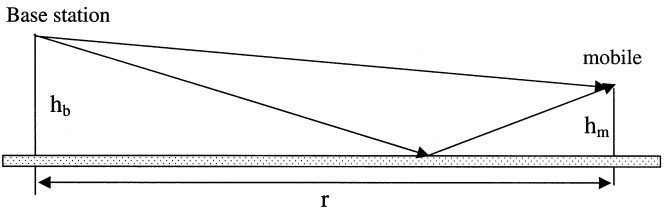


Figure 5.10 Radio propagation over a flat surface.

where r (in kilometers) is the distance between the base and mobile stations, as illustrated in Fig. 5.10. The values of A, B, C, and D are given in terms of the carrier frequency f (in MHz), the base station antenna height h_b (in meters), and the mobile station antenna height h_m (in meters) as

$$A = 69.55 + 26.16\log_{10} f - 13.82\log_{10} h_b - a(h_m) \tag{5.127a}$$

$$B = 44.9 - 6.55\log_{10} h_b \tag{5.127b}$$

$$C = 5.4 + 2\left(\log_{10}\frac{f}{28}\right)^2 \tag{5.127c}$$

$$D = 40.94 - 19.33\log_{10} f + 4.78\left(\log_{10} f\right)^2 \tag{5.127d}$$

where

$$a(h_m) = \begin{cases} 0.8 - 1.56\log_{10} f + \left(1.1\ \log_{10} f - 0.7\right)h_m & \text{for medium/small city} \\ 8.28\left[\log_{10}(1.54h_m)\right]^2 - 1.1 & \text{for } f \geq 200\,\text{MHz} \\ 3.2\left[\log_{10}(11.75h_m)\right]^2 - 4.97 & \text{for } f < 400\,\text{MHz for large city} \end{cases} \tag{5.128}$$

The following conditions must be satisfied before Eq. (5.127) is used: $150 < f < 1500$ MHz; $1 < r < 80$ km; $30 < h_b < 400$ m; $1 < h_m < 10$ m. Okumura's model has been found to be fairly good in urban and suburban areas but not as good in rural areas.

REFERENCES

1. David M. Pozar. *Microwave Engineering*; Addison-Wesley Publishing Company: New York, NY, 1990.
2. Kong, J.A. *Theory of Electromagnetic Waves*; Wiley: New York, 1975.
3. Sadiku, M.N.O. *Elements of Electromagnetics*, 3rd Ed.; Oxford University Press: New York, 2001; 621–623.
4. Sadiku, M.N.O. Wave propagation, In *The Electrical Engineering Handbook*; Dorf, R.C., Ed.; CRC Press: Boca Raton, FL, 1997; 925–937.
5. Blake, L.V. *Radar Range-Performance Analysis*; Artech House: Norwood, MA, 1986; 253–271.
6. Beckman, P.; Spizzichino, A. *The Scattering of Electromagnetic Waves from Random Surfaces*; Macmillan: New York, 1963.

© 2006 by Taylor & Francis Group, LLC

7. Smith, B.G. Geometrical shadowing of a random rough surface. IEEE Trans. Ant. Prog. **1967**, *15*, 668–671.
8. Okumura, Y. et al. Field strength and its variability in VHF and UHF land mobile service. Review of Electrical Communication Lab **Sept./Oct. 1969**, *16*, 825–873.
9. Feher, K. *Wireless Digital Communications*; Prentice-Hall: Upper Saddle River, NJ, 1995; 74–76.

© 2006 by Taylor & Francis Group, LLC

6
Transmission Lines

Andreas Weisshaar
Oregon State University
Corvallis, Oregon

6.1. INTRODUCTION

A transmission line is an electromagnetic guiding system for efficient point-to-point transmission of electric signals (information) and power. Since its earliest use in telegraphy by Samual Morse in the 1830s, transmission lines have been employed in various types of electrical systems covering a wide range of frequencies and applications. Examples of common transmission-line applications include TV cables, antenna feed lines, telephone cables, computer network cables, printed circuit boards, and power lines. A transmission line generally consists of two or more conductors embedded in a system of dielectric media. Figure 6.1 shows several examples of commonly used types of transmission lines composed of a set of parallel conductors.

The coaxial cable (Fig. 6.1a) consists of two concentric cylindrical conductors separated by a dielectric material, which is either air or an inert gas and spacers, or a foam-filler material such as polyethylene. Owing to their self-shielding property, coaxial cables are widely used throughout the radio frequency (RF) spectrum and in the microwave frequency range. Typical applications of coaxial cables include antenna feed lines, RF signal distribution networks (e.g., cable TV), interconnections between RF electronic equipment, as well as input cables to high-frequency precision measurement equipment such as oscilloscopes, spectrum analyzers, and network analyzers.

Another commonly used transmission-line type is the two-wire line illustrated in Fig. 6.1b. Typical examples of two-wire lines include overhead power and telephone lines and the flat twin-lead line as an inexpensive antenna lead-in line. Because the two-wire line is an open transmission-line structure, it is susceptible to electromagnetic interference. To reduce electromagnetic interference, the wires may be periodically twisted (twisted pair) and/or shielded. As a result, unshielded twisted pair (UTP) cables, for example, have become one of the most commonly used types of cable for high-speed local area networks inside buildings.

Figure 6.1c–e shows several examples of the important class of planar-type transmission lines. These types of transmission lines are used, for example, in printed circuit boards to interconnect components, as interconnects in electronic packaging, and as interconnects in integrated RF and microwave circuits on ceramic or semiconducting substrates. The microstrip illustrated in Fig. 6.1c consists of a conducting strip and a

© 2006 by Taylor & Francis Group, LLC

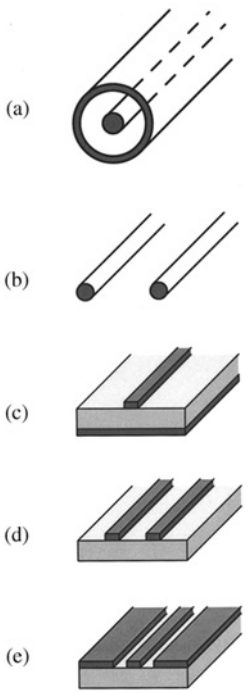


Figure 6.1 Examples of commonly used transmission lines: (a) coaxial cable, (b) two-wire line, (c) microstrip, (d) coplanar stripline, (e) coplanar waveguide.

conducting plane (ground plane) separated by a dielectric substrate. It is a widely used planar transmission line mainly because of its ease of fabrication and integration with devices and components. To connect a shunt component, however, through-holes are needed to provide access to the ground plane. On the other hand, in the coplanar stripline and coplanar waveguide (CPW) transmission lines (Fig. 6.1d and e) the conducting signal and ground strips are on the same side of the substrate. The single-sided conductor configuration eliminates the need for through-holes and is preferable for making connections to surface-mounted components.

In addition to their primary function as guiding system for signal and power transmission, another important application of transmission lines is to realize capacitive and inductive circuit elements, in particular at microwave frequencies ranging from a few gigahertz to tens of gigahertz. At these frequencies, lumped reactive elements become exceedingly small and difficult to realize and fabricate. On the other hand, transmission-line sections of appropriate lengths on the order of a quarter wavelength can be easily realized and integrated in planar transmission-line technology. Furthermore, transmission-line circuits are used in various configurations for impedance matching. The concept of functional transmission-line elements is further extended to realize a range of microwave passive components in planar transmission-line technology such as filters, couplers and power dividers [1].

This chapter on transmission lines provides a summary of the fundamental transmission-line theory and gives representative examples of important engineering applications. The following sections summarize the fundamental mathematical transmission-line equations and associated concepts, review the basic characteristics of transmission lines, present the transient response due to a step voltage or voltage pulse

© 2006 by Taylor & Francis Group, LLC

as well as the sinusoidal steady-state response of transmission lines, and give practical application examples and solution techniques. The chapter concludes with a brief summary of more advanced transmission-line concepts and gives a brief discussion of current technological developments and future directions.

6.2. BASIC TRANSMISSION-LINE CHARACTERISTICS

A transmission line is inherently a distributed system that supports propagating electromagnetic waves for signal transmission. One of the main characteristics of a transmission line is the delayed-time response due to the finite wave velocity.

The transmission characteristics of a transmission line can be rigorously determined by solving Maxwell's equations for the corresponding electromagnetic problem. For an "ideal" transmission line consisting of two parallel perfect conductors embedded in a homogeneous dielectric medium, the fundamental transmission mode is a transverse electromagnetic (TEM) wave, which is similar to a plane electromagnetic wave described in the previous chapter [2]. The electromagnetic field formulation for TEM waves on a transmission line can be converted to corresponding voltage and current circuit quantities by integrating the electric field between the conductors and the magnetic field around a conductor in a given plane transverse to the direction of wave propagation [3,4].

Alternatively, the transmission-line characteristics may be obtained by considering the transmission line directly as a distributed-parameter circuit in an extension of the traditional circuit theory [5]. The distributed circuit parameters, however, need to be determined from electromagnetic field theory. The distributed-circuit approach is followed in this chapter.

6.2.1. Transmission-line Parameters

A transmission line may be described in terms of the following distributed-circuit parameters, also called *line parameters*: the inductance parameter L (in H/m), which represents the series (loop) inductance per unit length of line, and the capacitance parameter C (in F/m), which is the shunt capacitance per unit length between the two conductors. To represent line losses, the resistance parameter R (in Ω/m) is defined for the series resistance per unit length due to the finite conductivity of both conductors, while the conductance parameter G (in S/m) gives the shunt conductance per unit length of line due to dielectric loss in the material surrounding the conductors.

The R, L, G, C transmission-line parameters can be derived in terms of the electric and magnetic field quantities by relating the corresponding stored energy and dissipated power. The resulting relationships are [1,2]

$$L = \frac{\mu}{|I|^2}\int_S \mathbf{H}\cdot\mathbf{H}^* ds \tag{6.1}$$

$$C = \frac{\epsilon'}{|V|^2}\int_S \mathbf{E}\cdot\mathbf{E}^* ds \tag{6.2}$$

$$R = \frac{R_s}{|I|^2}\int_{C_1+C_2} \mathbf{H}\cdot\mathbf{H}^* dl \tag{6.3}$$

$$G = \frac{\omega\epsilon'\tan\delta}{|V|^2}\int_S \mathbf{E}\cdot\mathbf{E}^* ds \tag{6.4}$$

© 2006 by Taylor & Francis Group, LLC

where **E** and **H** are the electric and magnetic field vectors in phasor form, "*" denotes complex conjugate operation, R_s is the surface resistance of the conductors,[†] ϵ' is the permittivity and $\tan\delta$ is the loss tangent of the dielectric material surrounding the conductors, and the line integration in Eq. (6.3) is along the contours enclosing the two conductor surfaces.

In general, the line parameters of a lossy transmission line are frequency dependent owing to the *skin effect* in the conductors and loss tangent of the dielectric medium.[‡] In the following, a lossless transmission line having constant L and C and zero R and G parameters is considered. This model represents a good first-order approximation for many practical transmission-line problems. The characteristics of lossy transmission lines are discussed in Sec. 6.4.

6.2.2. Transmission-line Equations for Lossless Lines

The fundamental equations that govern wave propagation on a lossless transmission line can be derived from an equivalent circuit representation for a short section of transmission line of length Δz illustrated in Fig. 6.2. A mathematically more rigorous derivation of the transmission-line equations is given in Ref. 5.

By considering the voltage drop across the series inductance $L\Delta z$ and current through the shunt capacitance $C\Delta z$, and taking $\Delta z \to 0$, the following fundamental transmission-line equations (also known as *telegrapher's equations*) are obtained.

$$\frac{\partial v(z,t)}{\partial z} = -L\frac{\partial i(z,t)}{\partial t} \tag{6.5}$$

$$\frac{\partial i(z,t)}{\partial z} = -C\frac{\partial v(z,t)}{\partial t} \tag{6.6}$$

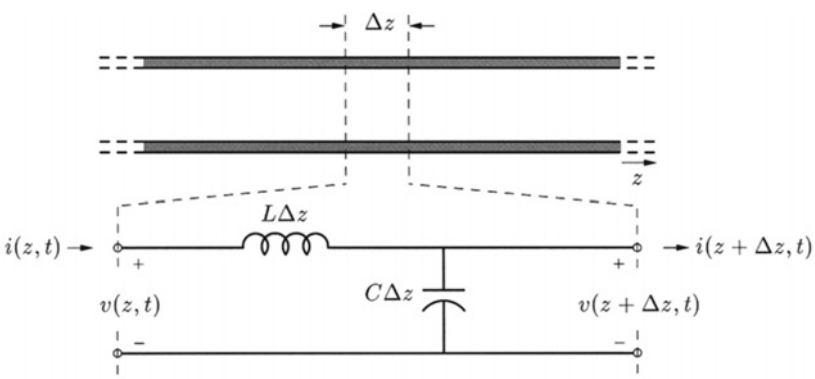


Figure 6.2 Schematic representation of a two-conductor transmission line and associated equivalent circuit model for a short section of lossless line.

[†]For a good conductor the surface resistance is $R_s = 1/\sigma\delta_s$, where the skin depth $\delta_s = 1/\sqrt{\pi f \mu\sigma}$ is assumed to be small compared to the cross-sectional dimensions of the conductor.

[‡]The skin effect describes the nonuniform current distribution inside the conductor caused by the time-varying magnetic flux within the conductor. As a result the resistance per unit length increases while the inductance per unit length decreases with increasing frequency. The loss tangent of the dielectric medium $\tan\delta = \epsilon''/\epsilon'$ typically results in an increase in shunt conductance with frequency, while the change in capacitance is negligible in most practical cases.

© 2006 by Taylor & Francis Group, LLC

The transmission-line equations, Eqs. (6.5) and (6.6), can be combined to obtain a one-dimensional wave equation for voltage

$$\frac{\partial^2 v(z,t)}{\partial z^2} = LC\frac{\partial^2 v(z,t)}{\partial t^2} \tag{6.7}$$

and likewise for current.

6.2.3. General Traveling-wave Solutions for Lossless Lines

The wave equation in Eq. (6.7) has the general solution

$$v(z,t) = v^+\left(t - \frac{z}{v_p}\right) + v^-\left(t + \frac{z}{v_p}\right) \tag{6.8}$$

where $v^+(t - z/v_p)$ corresponds to a wave traveling in the positive z direction, and $v^-(t + z/v_p)$ to a wave traveling in the negative z direction with constant velocity of propagation

$$v_p = \frac{1}{\sqrt{LC}} \tag{6.9}$$

Figure 6.3 illustrates the progression of a single traveling wave as function of position along the line and as function of time.

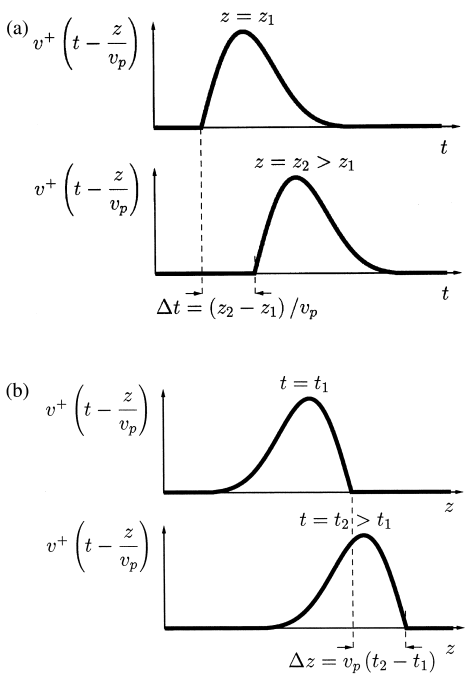


Figure 6.3 Illustration of the space and time variation for a general voltage wave $v^+(t - z/v_p)$: (a) variation in time and (b) variation in space.

© 2006 by Taylor & Francis Group, LLC

A corresponding solution for sinusoidal traveling waves is

$$\begin{aligned} v(z,t) &= v_0^+ \cos\left[\omega\left(t - \frac{z}{v_p}\right) + \phi^+\right] + v_0^- \cos\left[\omega\left(t + \frac{z}{v_p}\right) + \phi^-\right] \\ &= v_0^+ \cos(\omega t - \beta z + \phi^+) + v_0^- \cos(\omega t + \beta z + \phi^-) \end{aligned} \tag{6.10}$$

where

$$\beta = \frac{\omega}{v_p} = \frac{2\pi}{\lambda} \tag{6.11}$$

is the phase constant and $\lambda = v_p/f$ is the wavelength on the line. Since the spatial phase change βz depends on both the physical distance and the wavelength on the line, it is commonly expressed as *electrical distance* (or *electrical length*) θ with

$$\theta = \beta z = 2\pi \frac{z}{\lambda} \tag{6.12}$$

The corresponding wave solutions for current associated with voltage $v(z,t)$ in Eq. (6.8) are found with Eq. (6.5) or (6.6) as

$$i(z,t) = \frac{v^+(t - z/v_p)}{Z_0} - \frac{v^-(t + z/v_p)}{Z_0} \tag{6.13}$$

The parameter Z_0 is defined as the *characteristic impedance* of the transmission line and is given in terms of the line parameters by

$$Z_0 = \sqrt{\frac{L}{C}} \tag{6.14}$$

The characteristic impedance Z_0 specifies the ratio of voltage to current of a single traveling wave and, in general, is a function of both the conductor configuration (dimensions) and the electric and magnetic properties of the material surrounding the conductors. The negative sign in Eq. (6.13) for a wave traveling in the negative z direction accounts for the definition of positive current in the positive z direction.

As an example, consider the coaxial cable shown in Fig. 6.1a with inner conductor of diameter d, outer conductor of diameter D, and dielectric medium of dielectric constant ϵ_r. The associated distributed inductance and capacitance parameters are

$$L = \frac{\mu_0}{2\pi} \ln \frac{D}{d} \tag{6.15}$$

$$C = \frac{2\pi\epsilon_0\epsilon_r}{\ln(D/d)} \tag{6.16}$$

where $\mu_0 = 4\pi \times 10^{-7}$ H/m is the free-space permeability and $\epsilon_0 \approx 8.854 \times 10^{-12}$ F/m is the free-space permittivity. The characteristic impedance of the coaxial line is

$$Z_0 = \sqrt{\frac{L}{C}} = \frac{1}{2\pi}\sqrt{\frac{\mu_0}{\epsilon_0\epsilon_r}} \ln \frac{D}{d} = \frac{60}{\sqrt{\epsilon_r}} \ln \frac{D}{d} \qquad (\Omega) \tag{6.17}$$

© 2006 by Taylor & Francis Group, LLC

and the velocity of propagation is

$$v_p = \frac{1}{LC} = \frac{1}{\sqrt{\mu_0 \epsilon_0 \epsilon_r}} = \frac{c}{\sqrt{\epsilon_r}} \tag{6.18}$$

where $c \approx 30\,\text{cm/ns}$ is the velocity of propagation in free space.

In general, the velocity of propagation of a TEM wave on a lossless transmission line embedded in a homogeneous dielectric medium is independent of the geometry of the line and depends only on the material properties of the dielectric medium. The velocity of propagation is reduced from the free-space velocity c by the factor $1/\sqrt{\epsilon_r}$, which is also called the *velocity factor* and is typically given in percent.

For transmission lines with inhomogeneous or mixed dielectrics, such as the microstrip shown in Fig. 6.1c, the velocity of propagation depends on both the cross-sectional geometry of the line and the dielectric constants of the dielectric media. In this case, the electromagnetic wave propagating on the line is not strictly TEM, but for many practical applications can be approximated as a quasi-TEM wave. To extend Eq. (6.18) to transmission lines with mixed dielectrics, the inhomogeneous dielectric is replaced with a homogeneous dielectric of *effective dielectric constant* ϵ_{eff} giving the same capacitance per unit length as the actual structure. The effective dielectric constant is obtained as the ratio of the actual distributed capacitance C of the line to the capacitance of the same structure but with all dielectrics replaced with air:

$$\epsilon_{\text{eff}} = \frac{C}{C_{\text{air}}} \tag{6.19}$$

The velocity of propagation of the quasi-TEM wave can be expressed with Eq. (6.19) as

$$v_p = \frac{1}{\sqrt{\mu_0 \epsilon_0 \epsilon_{\text{eff}}}} = \frac{c}{\sqrt{\epsilon_{\text{eff}}}} \tag{6.20}$$

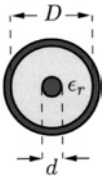
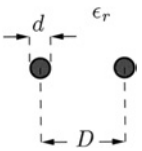
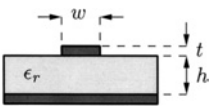
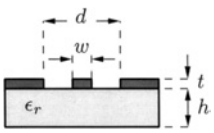
In general, the effective dielectric constant needs to be computed numerically; however, approximate closed-form expressions are available for many common transmission-line structures. As an example, a simple approximate closed-form expression for the effective dielectric constant of a microstrip of width w, substrate height h, and dielectric constant ϵ_r is given by [6]

$$\epsilon_{\text{eff}} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \frac{1}{\sqrt{1 + 10h/w}} \tag{6.21}$$

Various closed-form approximations of the transmission-line parameters for many common planar transmission lines have been developed and can be found in the literature including Refs. 6 and 7. Table 6.1 gives the transmission-line parameters in exact or approximate closed form for several common types of transmission lines (assuming no losses).

© 2006 by Taylor & Francis Group, LLC

Table 6.1 Transmission-line Parameters for Several Common Types of Transmission Lines

Transmission line	Parameters
 Coaxial line	$L = \frac{\mu_0}{2\pi}\ln(D/d)$ $C = \frac{2\pi\epsilon_0\epsilon_r}{\ln(D/d)}$ $Z_0 = \frac{1}{2\pi}\sqrt{\frac{\mu_0}{\epsilon_0\epsilon_r}}\ln(D/d)$ $\epsilon_{\text{eff}} = \epsilon_r$
 Two-wire line	$L = \frac{\mu_0}{\pi}\cosh^{-1}(D/d)$ $C = \frac{\pi\epsilon_0\epsilon_r}{\cosh^{-1}(D/d)}$ $Z_0 = \frac{1}{\pi}\sqrt{\frac{\mu_0}{\epsilon_0\epsilon_r}}\cosh^{-1}(D/d)$ $\epsilon_{\text{eff}} = \epsilon_r$
 Microstrip	$\epsilon_{\text{eff}} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2}\frac{1}{\sqrt{1 + 10h/w}}$ $Z_0 = \begin{cases} \frac{60}{\sqrt{\epsilon_{\text{eff}}}}\ln\left(\frac{8h}{w} + \frac{w}{4h}\right) & \text{for } w/h \le 1 \\ \frac{120\pi}{F\sqrt{\epsilon_{\text{eff}}}} & \text{for } w/h \ge 1 \end{cases}$ $F = w/h + 2.42 - 0.44h/w + (1 - h/w)^6$ $t \to 0$ [6]
 Coplanar waveguide	$\epsilon_{\text{eff}} = 1 + \frac{(\epsilon_r - 1)K(k_1')K(k')}{2K(k_1)K(k)}$ $k_1' = \sqrt{1 - k_1^2} = \frac{\sinh[\pi w/(4h)]}{\sinh[\pi d/(4h)]}$ $k' = \sqrt{1 - k^2} = \sqrt{(1 - (w/d)^2)}$ $Z_0 = \frac{30}{\sqrt{\epsilon_{\text{eff}}}}\frac{K(k')}{K(k)}$ $t \to 0$ [6] ($K(k)$ is the elliptical integral of the first kind)

6.3. TRANSIENT RESPONSE OF LOSSLESS TRANSMISSION LINES

A practical transmission line is of finite length and is necessarily terminated. Consider a transmission-line circuit consisting of a section of lossless transmission line that is connected to a source and terminated in a load, as illustrated in Fig. 6.4. The response of the transmission-line circuit depends on the transmission-line characteristics as well as the characteristics of the source and terminating load. The ideal transmission line of finite

© 2006 by Taylor & Francis Group, LLC

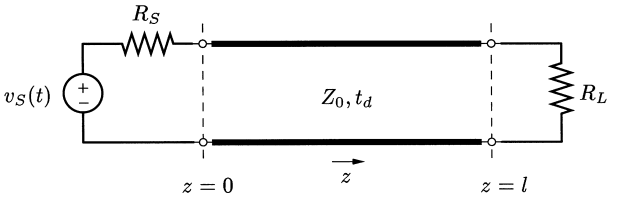


Figure 6.4 Lossless transmission line with resistive Thévénin equivalent source and resistive termination.

length is completely specified by the distributed L and C parameters and line length l, or, equivalently, by its characteristic impedance $Z_0 = \sqrt{L/C}$ and delay time

$$t_d = \frac{l}{v_p} = l\sqrt{LC} \tag{6.22}$$

of the line.* The termination imposes voltage and current boundary conditions at the end of the line, which may give rise to wave reflections.

6.3.1. Reflection Coefficient

When a traveling wave reaches the end of the transmission line, a reflected wave is generated unless the termination presents a load condition that is equal to the characteristic impedance of the line. The ratio of reflected voltage to incident voltage at the termination is defined as *voltage reflection coefficient* ρ, which for linear resistive terminations can be directly expressed in terms of the terminating resistance and the characteristic impedance of the line. The corresponding current reflection coefficient is given by $-\rho$. For the transmission-line circuit shown in Fig. 6.4 with resistive terminations, the voltage reflection coefficient at the termination with load resistance R_L is

$$\rho_L = \frac{R_L - Z_0}{R_L + Z_0} \tag{6.23}$$

Similarly, the voltage reflection coefficient at the source end with source resistance R_S is

$$\rho_S = \frac{R_S - Z_0}{R_S + Z_0} \tag{6.24}$$

The inverse relationship between reflection coefficient ρ_L and load resistance R_L follows directly from Eg. (6.23) and is

$$R_L = \frac{1 + \rho_L}{1 - \rho_L} Z_0 \tag{6.25}$$

*The specification in terms of characteristic impedance and delay time is used, for example, in the standard SPICE model for an ideal transmission line [8].

© 2006 by Taylor & Francis Group, LLC

It is seen from Eq. (6.23) or (6.24) that the reflection coefficient is positive for a termination resistance greater than the characteristic impedance, and it is negative for a termination resistance less than the characteristic impedance of the line. A termination resistance equal to the characteristic impedance produces no reflection ($\rho = 0$) and is called *matched termination*. For the special case of an open-circuit termination the voltage reflection coefficient is $\rho_{oc} = +1$, while for a short-circuit termination the voltage reflection coefficient is $\rho_{sc} = -1$.

6.3.2. Step Response

To illustrate the wave reflection process, the step-voltage response of an ideal transmission line connected to a Thévénin equivalent source and terminated in a resistive load, as shown in Fig. 6.4, is considered. The transient response for a step-voltage change with finite rise time can be obtained in a similar manner. The step-voltage response of a *lossy* transmission line with constant or frequency-dependent line parameters is more complex and can be determined using the Laplace transformation [5].

The source voltage $v_S(t)$ in the circuit in Fig. 6.4 is assumed to be a step-voltage given by

$$v_S(t) = V_0 U(t) \tag{6.26}$$

where

$$U(t) = \begin{cases} 1 & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \tag{6.27}$$

The transient response due to a rectangular pulse $v_{\text{pulse}}(t)$ of duration T can be obtained as the superposition of two step responses given as $v_{\text{pulse}}(t) = V_0 U(t) - V_0 U(t - T)$.

The step-voltage change launches a forward traveling wave at the input of the line at time $t = 0$. Assuming no initial charge or current on the line, this first wave component presents a resistive load to the generator that is equal to the characteristic impedance of the line. The voltage of the first traveling wave component is

$$v_1^+(z, t) = V_0 \frac{Z_0}{R_S + Z_0} U\left(t - \frac{z}{v_p}\right) = V_1^+ U\left(t - \frac{z}{v_p}\right) \tag{6.28}$$

where v_p is the velocity of propagation on the line. For a nonzero reflection coefficient ρ_L at the termination, a reflected wave is generated when the first traveling wave arrives at the termination at time $t = t_d = l/v_p$. If the reflection coefficients at both the source and the termination are nonzero, an infinite succession of reflected waves results. The total voltage

© 2006 by Taylor & Francis Group, LLC

response on the line is the superposition of all traveling-wave components and is given by

$$\begin{aligned} v(z,t) = \frac{Z_0}{R_S + Z_0} V_0 \Bigg[& U\left(t - \frac{z}{v_p}\right) + \rho_L U\left(t - 2t_d + \frac{z}{v_p}\right) \\ & + \rho_S \rho_L U\left(t - 2t_d - \frac{z}{v_p}\right) + \rho_S \rho_L^2 U\left(t - 4t_d + \frac{z}{v_p}\right) \\ & + \rho_S^2 \rho_L^2 U\left(t - 4t_d - \frac{z}{v_p}\right) + \rho_S^2 \rho_L^3 U\left(t - 6t_d + \frac{z}{v_p}\right) \\ & + \cdots \Bigg] \end{aligned} \tag{6.29}$$

Similarly, the total current on the line is given by

$$\begin{aligned} i(z,t) = \frac{V_0}{R_S + Z_0} \Bigg[& U\left(t - \frac{z}{v_p}\right) - \rho_L U\left(t - 2t_d + \frac{z}{v_p}\right) \\ & + \rho_S \rho_L U\left(t - 2t_d - \frac{z}{v_p}\right) - \rho_S \rho_L^2 U\left(t - 4t_d + \frac{z}{v_p}\right) \\ & + \rho_S^2 \rho_L^2 U\left(t - 4t_d - \frac{z}{v_p}\right) - \rho_S^2 \rho_L^3 U\left(t - 6t_d + \frac{z}{v_p}\right) \\ & + \cdots \Bigg] \end{aligned} \tag{6.30}$$

The reflected wave components on the lossless transmission line are successively delayed copies of the first traveling-wave component with amplitudes appropriately adjusted by the reflection coefficients. Equations (6.29) and (6.30) show that at any given time and location on the line only a finite number of wave components have been generated. For example, for $t = 3t_d$ three wave components exist at the input of the line (at $z = 0$) and four wave components exist at the load (at $z = l$).

Unless both reflection coefficients have unity magnitudes, the amplitudes of the successive wave components become progressively smaller in magnitude and the infinite summations in Eqs. (6.29) and (6.30) converge to the dc values for $t \to \infty$. The steady-state (dc) voltage V_∞ is obtained by summing the amplitudes of all traveling-wave components for $t \to \infty$.

$$\begin{aligned} V_\infty = v(z, t \to \infty) &= \frac{Z_0}{R_S + Z_0} V_0 \{1 + \rho_L + \rho_S \rho_L + \rho_S \rho_L^2 + \rho_S^2 \rho_L^2 + \cdots\} \\ &= \frac{Z_0}{R_S + Z_0} V_0 \frac{1 + \rho_L}{1 - \rho_S \rho_L} \end{aligned} \tag{6.31}$$

The steady-state voltage can also be directly obtained as the dc voltage drop across the load after removing the lossless line, that is

$$V_\infty = \frac{R_L}{R_S + R_L} V_0 \tag{6.32}$$

© 2006 by Taylor & Francis Group, LLC

The steady-state current is

$$I_\infty = \frac{V_0}{R_S + R_L} \tag{6.33}$$

6.3.3. Lattice Diagram

The *lattice diagram* (also called *bounce* or *reflection diagram*) provides a convenient graphical means for keeping track of the multiple wave reflections on the line. The general lattice diagram is illustrated in Fig. 6.5. Each wave component is represented by a sloped line segment that shows the time elapsed after the initial voltage change at the source as a function of distance z on the line. For bookkeeping purposes, the value of the voltage amplitude of each wave component is commonly written above the corresponding line segment and the value of the accompanying current is added below. Starting with voltage $V_1^+ = V_0 Z_0/(R_S + Z_0)$ of the first wave component, the voltage amplitude of each successive wave is obtained from the voltage of the preceding wave by multiplication with the appropriate reflection coefficient ρ_L or ρ_S in accordance with Eq. (6.29). Successive current values are obtained by multiplication with $-\rho_L$ or $-\rho_S$, as shown in Eq. (6.30).

The lattice diagram may be conveniently used to determine the voltage and current distributions along the transmission line at any given time or to find the time response at any given position. The variation of voltage and current as a function of time at a given position $z = z_1$ is found from the intersection of the vertical line through z_1 and the sloped line segments representing the wave components. Figure 6.5 shows the first five wave intersection times at position z_1 marked as t_1, t_2, t_3, t_4, and t_5, respectively. At each

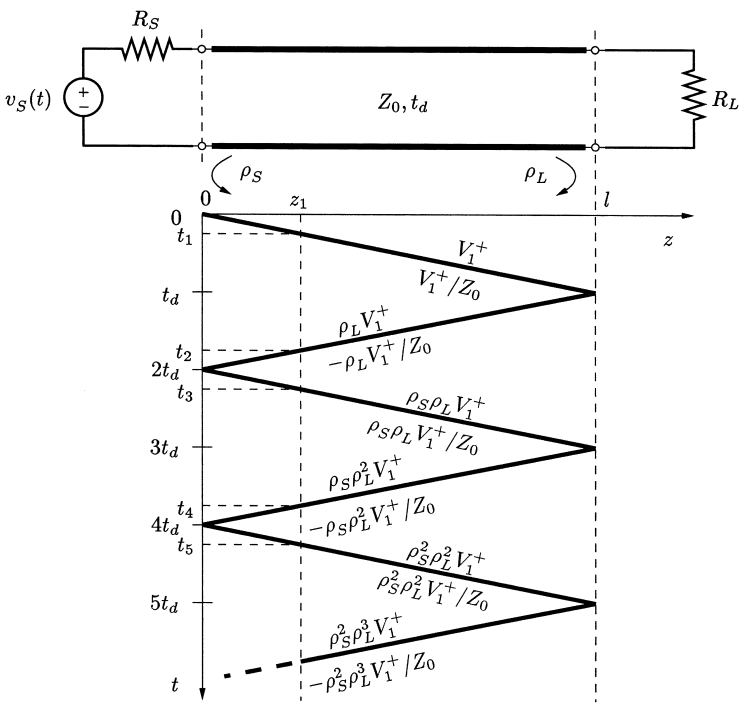


Figure 6.5 Lattice diagram for a lossless transmission line with unmatched terminations.

© 2006 by Taylor & Francis Group, LLC

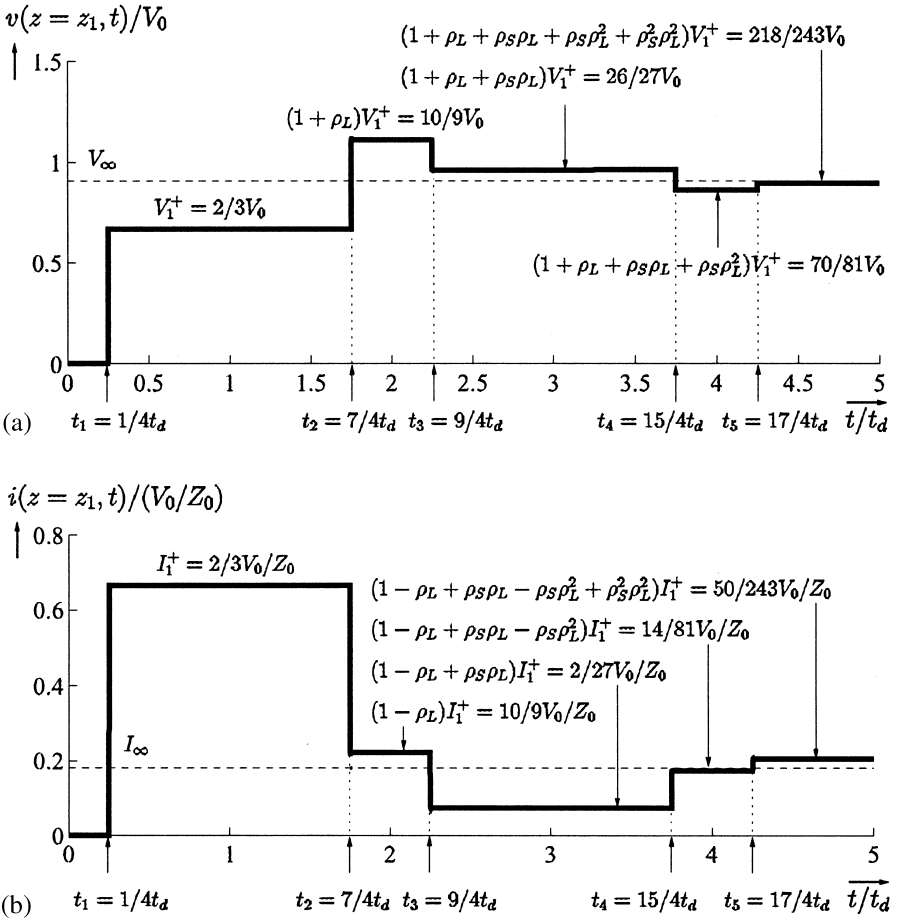

$v(z = z_1, t)/V_0$
$(1 + \rho_L + \rho_S\rho_L + \rho_S\rho_L^2 + \rho_S^2\rho_L^2)V_1^+ = 218/243V_0$
$(1 + \rho_L + \rho_S\rho_L)V_1^+ = 26/27V_0$
$(1 + \rho_L)V_1^+ = 10/9V_0$
V_∞
$V_1^+ = 2/3V_0$
$(1 + \rho_L + \rho_S\rho_L + \rho_S\rho_L^2)V_1^+ = 70/81V_0$
1.5
1
0.5
0
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
(a) $t_1 = 1/4t_d$ $t_2 = 7/4t_d$ $t_3 = 9/4t_d$ $t_4 = 15/4t_d$ $t_5 = 17/4t_d$ t/t_d
$i(z = z_1, t)/(V_0/Z_0)$
$I_1^+ = 2/3V_0/Z_0$
$(1 - \rho_L + \rho_S\rho_L - \rho_S\rho_L^2 + \rho_S^2\rho_L^2)I_1^+ = 50/243V_0/Z_0$
$(1 - \rho_L + \rho_S\rho_L - \rho_S\rho_L^2)I_1^+ = 14/81V_0/Z_0$
$(1 - \rho_L + \rho_S\rho_L)I_1^+ = 2/27V_0/Z_0$
$(1 - \rho_L)I_1^+ = 10/9V_0/Z_0$
I_∞
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5
(b) $t_1 = 1/4t_d$ $t_2 = 7/4t_d$ $t_3 = 9/4t_d$ $t_4 = 15/4t_d$ $t_5 = 17/4t_d$ t/t_d

Figure 6.6 Step response of a lossless transmission line at $z = z_1 = l/4$ for $R_S = Z_0/2$ and $R_L = 5Z_0$; (a) voltage response, (b) current response.

intersection time, the total voltage and current change by the amplitudes specified for the intersecting wave component. The corresponding transient response for voltage and current with $R_S = Z_0/2$ and $R_L = 5Z_0$ corresponding to reflection coefficients $\rho_S = -1/3$ and $\rho_L = 2/3$, respectively, is shown in Fig. 6.6. The transient response converges to the steady-state $V_\infty = 10/11\, V_0$ and $I_\infty = 2/11(V_0/Z_0)$, as indicated in Fig. 6.6.

6.3.4. Applications

In many practical applications, one or both ends of a transmission line are matched to avoid multiple reflections. If the source and/or the receiver do not provide a match, multiple reflections can be avoided by adding an appropriate resistor at the input of the line (source termination) or at the end of the line (end termination) [9,10]. Multiple reflections on the line may lead to signal distortion including a slow voltage buildup or signal overshoot and ringing.

© 2006 by Taylor & Francis Group, LLC

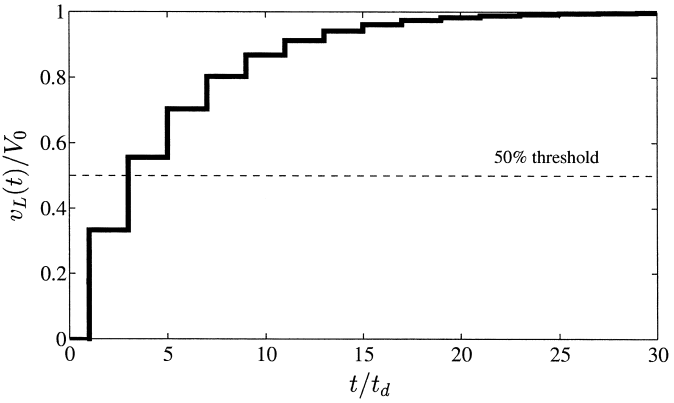


Figure 6.7 Step-voltage response at the termination of an open-circuited lossless transmission line with $R_S = 5Z_0$ ($\rho_S = 2/3$).

Over- and Under-driven Transmission Lines

In high-speed digital systems, the input of a receiver circuit typically presents a load to a transmission line that is approximately an open circuit (unterminated). The step-voltage response of an unterminated transmission line may exhibit a considerably different behavior depending on the source resistance.

If the source resistance is larger than the characteristic impedance of the line, the voltage across the load will build up monotonically to its final value since both reflection coefficients are positive. This condition is referred to as an *underdriven* transmission line. The buildup time to reach a sufficiently converged voltage may correspond to many round-trip times if the reflection coefficient at the source is close to $+1$ (and $\rho_L = \rho_{oc} = +1$), as illustrated in Fig. 6.7. As a result, the effective signal delay may be several times longer than the delay time of the line.

If the source resistance is smaller than the characteristic impedance of the line, the initial voltage at the unterminated end will exceed the final value (overshoot). Since the source reflection coefficient is negative and the load reflection coefficient is positive, the voltage response will exhibit ringing as the voltage converges to its final value. This condition is referred to as an *overdriven* transmission line. It may take many round-trip times to reach a sufficiently converged voltage (long settling time) if the reflection coefficient at the source is close to -1 (and $\rho_L = \rho_{oc} = +1$), as illustrated in Fig. 6.8. An overdriven line can produce excessive noise and cause intersymbol interference.

Transmission-line Junctions

Wave reflections occur also at the junction of two tandem-connected transmission lines having different characteristic impedances. This situation, illustrated in Fig. 6.9a, is often encountered in practice. For an incident wave on line 1 with characteristic impedance $Z_{0,1}$, the second line with characteristic impedance $Z_{0,2}$ presents a load resistance to line 1 that is equal to $Z_{0,2}$. At the junction, a reflected wave is generated on line 1 with voltage reflection coefficient ρ_{11} given by

$$\rho_{11} = \frac{Z_{0,2} - Z_{0,1}}{Z_{0,2} + Z_{0,1}} \tag{6.34}$$

© 2006 by Taylor & Francis Group, LLC

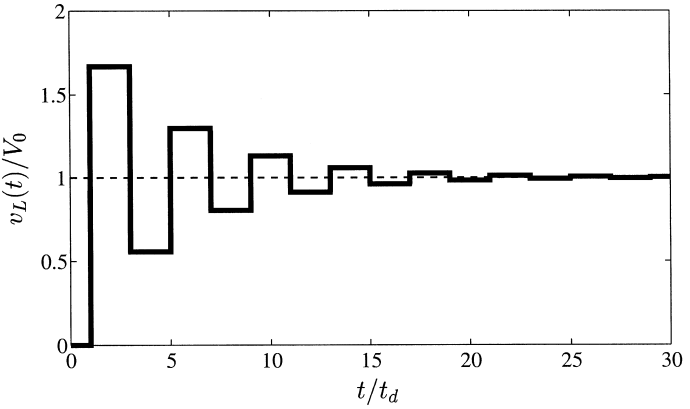


Figure 6.8 Step-voltage response at the termination of an open-circuited lossless transmission line with $R_S = Z_0/5$ ($\rho_S = -2/3$).

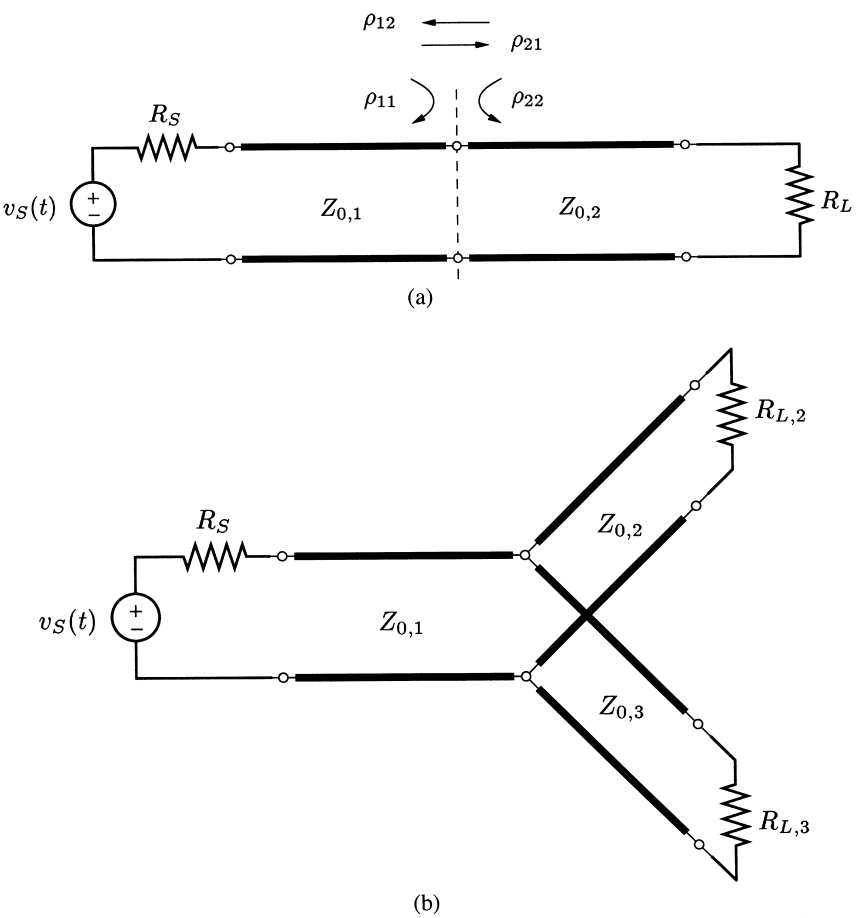


Figure 6.9 Junction between transmission lines: (a) two tandem-connected lines and (b) three parallel-connected lines.

© 2006 by Taylor & Francis Group, LLC

In addition, a wave is launched on the second line departing from the junction. The voltage amplitude of the transmitted wave is the sum of the voltage amplitudes of the incident and reflected waves on line 1. The ratio of the voltage amplitudes of the transmitted wave on line 2 to the incident wave on line 1 is defined as the *voltage transmission coefficient* ρ_{21} and is given by

$$\rho_{21} = 1 + \rho_{11} = \frac{2Z_{0,2}}{Z_{0,1} + Z_{0,2}} \tag{6.35}$$

Similarly, for an incident wave from line 2, the reflection coefficient ρ_{22} at the junction is

$$\rho_{22} = \frac{Z_{0,1} - Z_{0,2}}{Z_{0,1} + Z_{0,2}} = -\rho_{11} \tag{6.36}$$

The voltage transmission coefficient ρ_{12} for a wave incident from line 2 and transmitted into line 1 is

$$\rho_{12} = 1 + \rho_{22} = \frac{2Z_{0,1}}{Z_{0,1} + Z_{0,2}} \tag{6.37}$$

If in addition lumped elements are connected at the junction or the transmission lines are connected through a resistive network, the reflection and transmission coefficients will change, and in general, $\rho_{ij} \leq 1 + \rho_{jj}$ [5].

For a parallel connection of multiple lines at a common junction, as illustrated in Fig. 6.9b, the effective load resistance is obtained as the parallel combination of the characteristic impedances of all lines except for the line carrying the incident wave. The reflection and transmission coefficients are then determined as for tandem connected lines [5].

The wave reflection and transmission process for tandem and multiple parallel-connected lines can be represented graphically with a lattice diagram for each line. The complexity, however, is significantly increased over the single line case, in particular if multiple reflections exist.

Reactive Terminations

In various transmission-line applications, the load is not purely resistive but has a reactive component. Examples of reactive loads include the capacitive input of a CMOS gate, pad capacitance, bond-wire inductance, as well as the reactance of vias, package pins, and connectors [9,10]. When a transmission line is terminated in a reactive element, the reflected waveform will not have the same shape as the incident wave, i.e., the reflection coefficient will not be a constant but be varying with time. For example, consider the step response of a transmission line that is terminated in an uncharged capacitor C_L. When the incident wave reaches the termination, the initial response is that of a short circuit, and the response after the capacitor is fully charged is an open circuit. Assuming the source end is matched to avoid multiple reflections, the incident step-voltage wave is $v_1^+(t) = V_0/2U(t - z/v_p)$. The voltage across the capacitor changes exponentially from the initial voltage $v_{\text{cap}} = 0$ (short circuit) at time $t = t_d$ to the final voltage

© 2006 by Taylor & Francis Group, LLC

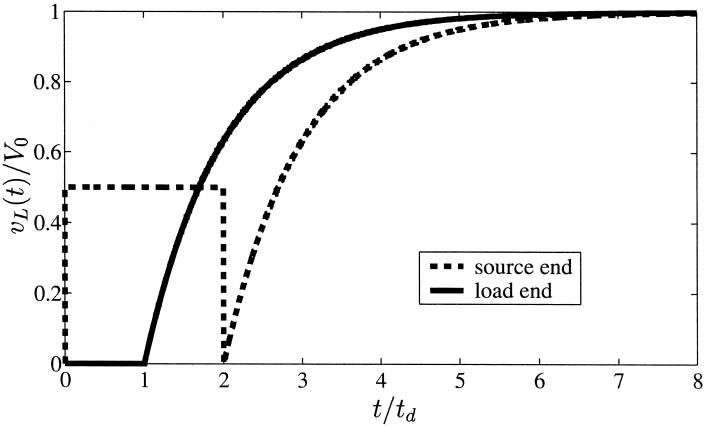


Figure 6.10 Step-voltage response of a transmission line that is matched at the source and terminated in a capacitor C_L with time constant $\tau = Z_0 C_L = t_d$.

$v_{\text{cap}}(t \to \infty) = V_0$ (open circuit) as

$$v_{\text{cap}}(t) = V_0\left[1 - e^{-(t-t_d)/\tau}\right]U(t - t_d) \tag{6.38}$$

with time constant

$$\tau = Z_0 C_L \tag{6.39}$$

where Z_0 is the characteristic impedance of the line. Figure 6.10 shows the step-voltage response across the capacitor and at the source end of the line for $\tau = t_d$.

If the termination consists of a parallel combination of a capacitor C_L and a resistor R_L, the time constant is obtained as the product of C_L and the parallel combination of R_L and characteristic impedance Z_0. For a purely inductive termination L_L, the initial response is an open circuit and the final response is a short circuit. The corresponding time constant is $\tau = L_L/Z_0$.

In the general case of reactive terminations with multiple reflections or with more complicated source voltages, the boundary conditions for the reactive termination are expressed in terms of a differential equation. The transient response can then be determined mathematically, for example, using the Laplace transformation [11].

Nonlinear Terminations

For a nonlinear load or source, the reflected voltage and subsequently the reflection coefficient are a function of the cumulative voltage and current at the termination including the contribution of the reflected wave to be determined. Hence, the reflection coefficient for a nonlinear termination cannot be found from only the termination characteristics and the characteristic impedance of the line. The step-voltage response for each reflection instance can be determined by matching the *I–V* characteristics of the termination and the cumulative voltage and current characteristics at the end of the transmission line. This solution process can be constructed using a graphical technique known as the *Bergeron method* [5,12] and can be implemented in a computer program.

© 2006 by Taylor & Francis Group, LLC

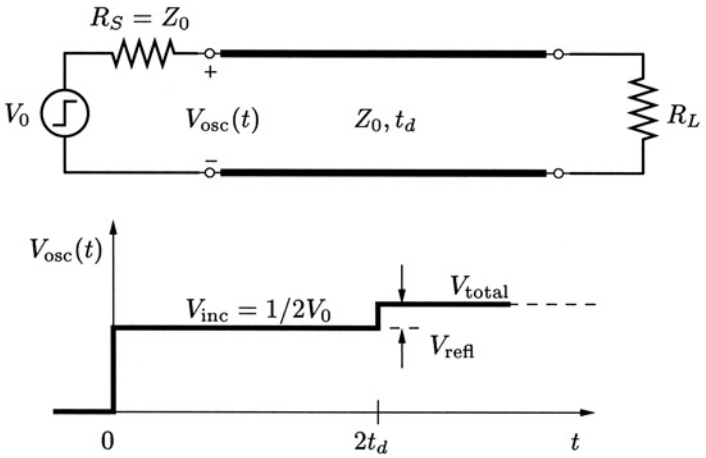


Figure 6.11 Illustration of the basic principle of time-domain reflectometry (TDR).

Time-Domain Reflectometry

Time-domain reflectometry (TDR) is a measurement technique that utilizes the information contained in the reflected waveform and observed at the source end to test, characterize, and model a transmission-line circuit. The basic TDR principle is illustrated in Fig. 6.11. A TDR instrument typically consists of a precision step-voltage generator with a known source (reference) impedance to launch a step wave on the transmission-line circuit under test and a high impedance probe and oscilloscope to sample and display the voltage waveform at the source end. The source end is generally well matched to establish a reflection-free reference. The voltage at the input changes from the initial incident voltage when a reflected wave generated at an impedance discontinuity such as a change in line impedance, a line break, an unwanted parasitic reactance, or an unmatched termination reaches the source end of the transmission line-circuit.

The time elapsed between the initial launch of the step wave and the observation of the reflected wave at the input corresponds to the round-trip delay $2t_d$ from the input to the location of the impedance mismatch and back. The round-trip delay time can be converted to find the distance from the input of the line to the location of the impedance discontinuity if the propagation velocity is known. The capability of measuring distance is used in TDR cable testers to locate faults in cables. This measurement approach is particularly useful for testing long, inaccessible lines such as underground or undersea electrical cables.

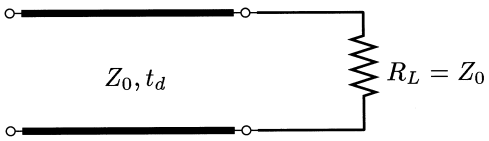
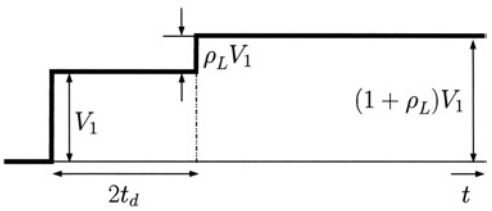
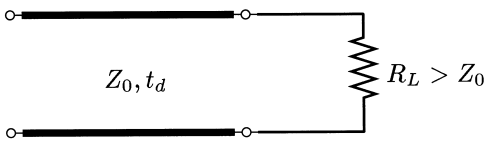
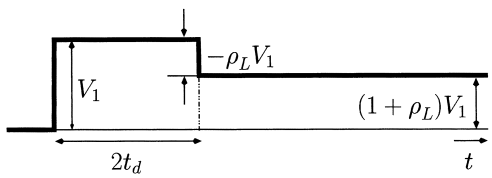
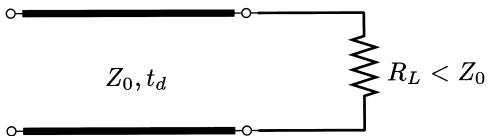
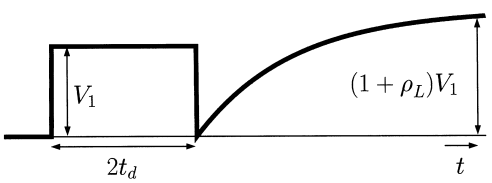
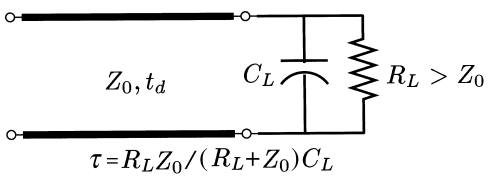
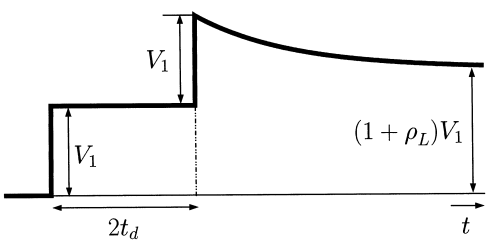
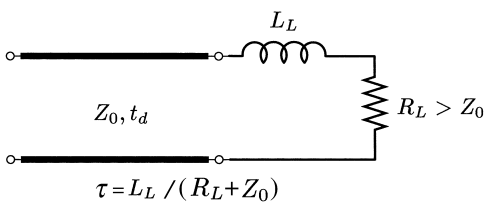
The reflected waveform observed at the input also provides information on the type of discontinuity and the amount of impedance change. Table 6.2 shows the TDR response for several common transmission-line discontinuities. As an example, the load resistance in the circuit in Fig. 6.11 is extracted from the incident and reflected or total voltage observed at the input as

$$R_L = Z_0 \frac{1+\rho}{1-\rho} = Z_0 \frac{V_{\text{total}}}{2V_{\text{incident}} - V_{\text{total}}} \tag{6.40}$$

where $\rho = V_{\text{reflected}}/V_{\text{incident}} = (R_L - Z_0)/(R_L + Z_0)$ and $V_{\text{total}} = V_{\text{incident}} + V_{\text{reflected}}$.

© 2006 by Taylor & Francis Group, LLC

Table 6.2 TDR Responses for Typical Transmission-line Discontinuities.

TDR response	Circuit
	
	
	
	
	

The TDR principle can be used to profile impedance changes along a transmission line circuit such as a trace on a printed-circuit board. In general, the effects of multiple reflections arising from the impedance mismatches along the line need to be included to extract the impedance profile. If the mismatches are small, higher-order reflections can be ignored and the same extraction approach as for a single impedance discontinuity can be applied for each discontinuity. The resolution of two closely spaced discontinuities, however, is limited by the rise time of step voltage and the overall rise time of the TDR system. Further information on using time-domain reflectometry for analyzing and modeling transmission-line systems is given e.g. in Refs. 10,11,13–15.

© 2006 by Taylor & Francis Group, LLC

6.4. SINUSOIDAL STEADY-STATE RESPONSE OF TRANSMISSION LINES

The steady-state response of a transmission line to a sinusoidal excitation of a given frequency serves as the fundamental solution for many practical transmission-line applications including radio and television broadcast and transmission-line circuits operating at microwave frequencies. The frequency-domain information also provides physical insight into the signal propagation on the transmission line. In particular, transmission-line losses and any frequency dependence in the R, L, G, C line parameters can be readily taken into account in the frequency-domain analysis of transmission lines. The time-domain response of a transmission-line circuit to an arbitrary time-varying excitation can then be obtained from the frequency-domain solution by applying the concepts of Fourier analysis [16].

As in standard circuit analysis, the time-harmonic voltage and current on the transmission line are conveniently expressed in phasor form using Euler's identity $e^{j\theta} = \cos\theta + j\sin\theta$. For a cosine reference, the relations between the voltage and current phasors, $V(z)$ and $I(z)$, and the time-harmonic space–time-dependent quantities, $v(z, t)$ and $i(z, t)$, are

$$v(z, t) = \mathrm{Re}\{V(z)e^{j\omega t}\} \tag{6.41}$$

$$i(z, t) = \mathrm{Re}\{I(z)e^{j\omega t}\} \tag{6.42}$$

The voltage and current phasors are functions of position z on the transmission line and are in general complex.

6.4.1. Characteristics of Lossy Transmission Lines

The transmission-line equations, (general telegrapher's equations) in phasor form for a general lossy transmission line can be derived directly from the equivalent circuit for a short line section of length $\Delta z \to 0$ shown in Fig. 6.12. They are

$$-\frac{dV(z)}{dz} = (R + j\omega L)I(z) \tag{6.43}$$

$$-\frac{dI(z)}{dz} = (G + j\omega C)V(z) \tag{6.44}$$

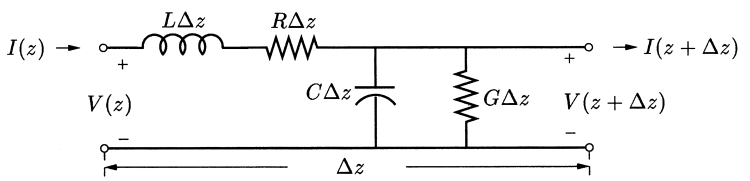


Figure 6.12 Equivalent circuit model for a short section of lossy transmission line of length Δz with R, L, G, C line parameters.

© 2006 by Taylor & Francis Group, LLC

The transmission-line equations, Eqs. (6.43) and (6.44) can be combined to the complex wave equation for voltage (and likewise for current)

$$\frac{d^2V(z)}{dz^2} = (R + j\omega L)(G + j\omega C)V(z) = \gamma^2 V(z) \tag{6.45}$$

The general solution of Eq. (6.45) is

$$V(z) = V^+(z) + V^-(z) = V_0^+ e^{-\gamma z} + V_0^- e^{+\gamma z} \tag{6.46}$$

where γ is the *propagation constant* of the transmission line and is given by

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \tag{6.47}$$

and $V_0^+ = |V_0^+|e^{j\phi^+}$ and $V_0^- = |V_0^-|e^{j\phi^-}$ are complex constants. The real time-harmonic voltage waveforms $v(z, t)$ corresponding to phasor $V(z)$ are obtained with Eq. (6.41) as

$$\begin{aligned} v(z,t) &= v^+(z,t) + v^-(z,t) \\ &= |V_0^+|e^{-\alpha z}\cos(\omega t - \beta z + \phi^+) + |V_0^-|e^{\alpha z}\cos(\omega t + \beta z + \phi^-) \end{aligned} \tag{6.48}$$

and are illustrated in Fig. 6.13.

The real part α of the propagation constant in Eq. (6.47) is known as the *attenuation constant* measured in nepers per unit length (Np/m) and gives the rate of exponential attenuation of the voltage and current amplitudes of a traveling wave.* The imaginary part of γ is the *phase constant* $\beta = 2\pi/\lambda$ measured in radians per unit length (rad/m), as in the lossless line case. The corresponding *phase velocity* of the time-harmonic wave is given by

$$v_p = \frac{\omega}{\beta} \tag{6.49}$$

which depends in general on frequency. Transmission lines with frequency-dependent phase velocity are called *dispersive* lines. Dispersive transmission lines can lead to signal distortion, in particular for broadband signals.

The current phasor $I(z)$ associated with voltage $V(z)$ in Eq. (6.46) is found with Eq. (6.43) as

$$I(z) = \frac{V^+}{Z_0}e^{-\gamma z} - \frac{V^-}{Z_0}e^{+\gamma z} \tag{6.50}$$

*The amplitude attenuation of a traveling wave $V^+(z) = V_0^+ e^{-\gamma z} = V_0^+ e^{-\alpha z}e^{-j\beta z}$ over a distance l can be expressed in logarithmic form as $\ln|V^+(z)/V^+(z+l)| = \alpha l$ (nepers). To convert from the attenuation measured in nepers to the logarithmic measure $20\log_{10}|V^+(z)/V^+(z+l)|$ in dB, the attenuation in nepers is multiplied by $20\log_{10} e \approx 8.686$ (1 Np corresponds to about 8.686 dB). For coaxial cables the attenuation constant is typically specified in units of dB/100 ft. The conversion to Np/m is 1 dB/100 ft $\approx$ 0.0038 Np/m.

© 2006 by Taylor & Francis Group, LLC

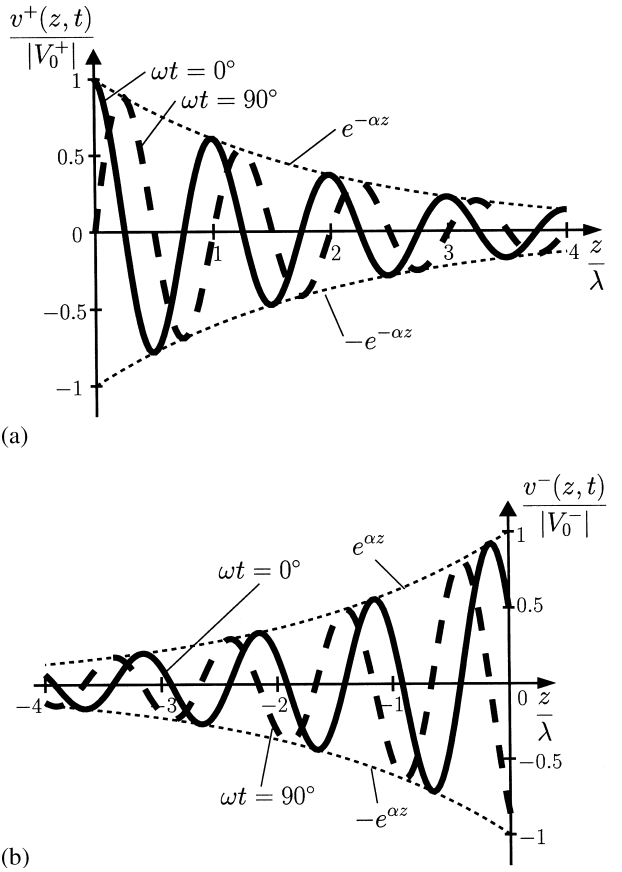


Figure 6.13 Illustration of a traveling wave on a lossy transmission line: (a) wave traveling in $+z$ direction with $\phi^+ = 0$ and $\alpha = 1/(2\lambda)$ and (b) wave traveling in $-z$ direction with $\phi^- = 60°$ and $\alpha = 1/(2\lambda)$.

The quantity Z_0 is defined as the characteristic impedance of the transmission line and is given in terms of the line parameters by

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{6.51}$$

As seen from Eq. (6.51), the characteristic impedance is in general complex and frequency dependent.

The inverse expressions relating the R, L, G, C line parameters to the characteristic impedance and propagation constant of a transmission line are found from Eqs. (6.47) and (6.51) as

$$R + j\omega L = \gamma Z_0 \tag{6.52}$$

$$G + j\omega C = \gamma / Z_0 \tag{6.53}$$

© 2006 by Taylor & Francis Group, LLC

These inverse relationships are particularly useful for extracting the line parameters from experimentally determined data for characteristic impedance and propagation constant.

Special Cases

For a lossless line with $R=0$ and $G=0$, the propagation constant is $\gamma = j\omega\sqrt{LC}$. The attenuation constant α is zero and the phase velocity is $v_p = \omega/\beta = 1/\sqrt{LC}$. The characteristic impedance of a lossless line is $Z_0 = \sqrt{L/C}$, as in Eq. (6.14).

In general, for a lossy transmission line both the attenuation constant and the phase velocity are frequency dependent, which can give rise to signal distortion.* However, in many practical applications the losses along the transmission line are small. For a low loss line with $R \ll \omega L$ and $G \ll \omega C$, useful approximate expressions can be derived for the characteristic impedance Z_0 and propagation constant γ as

$$Z_0 \approx \sqrt{\frac{L}{C}}\left[1 - j\frac{1}{2\omega}\left(\frac{R}{L} - \frac{G}{C}\right)\right] \tag{6.54}$$

and

$$\gamma \approx \frac{R}{2}\sqrt{\frac{C}{L}} + \frac{G}{2}\sqrt{\frac{L}{C}} + j\omega\sqrt{LC} \tag{6.55}$$

The low-loss conditions $R \ll \omega L$ and $G \ll \omega C$ are more easily satisfied at higher frequencies.

6.4.2. Terminated Transmission lines

If a transmission line is terminated with a load impedance that is different from the characteristic impedance of the line, the total time-harmonic voltage and current on the line will consist of two wave components traveling in opposite directions, as given by the general phasor expressions in Eqs. (6.46) and (6.50). The presence of the two wave components gives rise to standing waves on the line and affects the line's input impedance.

Impedance Transformation

Figure 6.14 shows a transmission line of finite length terminated with load impedance Z_L. In the steady-state analysis of transmission-line circuits it is expedient to measure distance on the line from the termination with known load impedance. The distance on the line from the termination is given by z'. The line voltage and current at distance z' from the

*For the special case of a line satisfying the condition $R/L = G/C$, the characteristic impedance $Z_0 = \sqrt{L/C}$, the attenuation constant $\alpha = R/\sqrt{L/C}$, and the phase velocity $v_p = 1/\sqrt{LC}$ are frequency independent. This type of line is called a *distortionless line*. Except for a constant signal attenuation, a distortionless line behaves like a lossless line.

© 2006 by Taylor & Francis Group, LLC

termination can be related to voltage $V_L = V(z' = 0)$ and current $I_L = I(z' = 0)$ at the termination as

$$V(z') = V_L \cosh \gamma z' + I_L Z_0 \sinh \gamma z' \tag{6.56}$$

$$I(z') = V_L\left(\frac{1}{Z_0}\right) \sinh \gamma z' + I_L \cosh \gamma z' \tag{6.57}$$

where $V_L/I_L = Z_L$. These voltage and current transformations between the input and output of a transmission line of length z' can be conveniently expressed in $ABCD$ matrix form as*

$$\begin{bmatrix} V(z') \\ I(z') \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V(0) \\ I(0) \end{bmatrix} = \begin{bmatrix} \cosh(\gamma z') & Z_0 \sinh(\gamma z') \\ (1/Z_0)\sinh(\gamma z') & \cosh(\gamma z') \end{bmatrix} \begin{bmatrix} V(0) \\ I(0) \end{bmatrix} \tag{6.58}$$

The ratio $V(z')/I(z')$ defines the input impedance $Z_{\text{in}}(z')$ at distance z' looking toward the load. The input impedance for a general lossy line with characteristic impedance Z_0 and terminated with load impedance Z_L is

$$Z_{\text{in}}(z') = \frac{V(z')}{I(z')} = Z_0 \frac{Z_L + Z_0 \tanh \gamma z'}{Z_0 + Z_L \tanh \gamma z'} \tag{6.60}$$

It is seen from Eq. (6.60) that for a line terminated in its characteristic impedance ($Z_L = Z_0$), the input impedance is identical to the characteristic impedance, independent of distance z'. This property serves as an alternate definition of the characteristic impedance of a line and can be applied to experimentally determine the characteristic impedance of a given line.

The input impedance of a transmission line can be used advantageously to determine the voltage and current at the input terminals of a transmission-line circuit as well as the average power delivered by the source and ultimately the average power dissipated in the load. Figure 6.15 shows the equivalent circuit at the input (source end) for the transmission-line circuit in Fig. 6.14. The input voltage V_{in} and current I_{in} are easily

*The $ABCD$ matrix is a common representation for two-port networks and is particularly useful for cascade connections of two or more two-port networks. The overall voltage and current transformations for cascaded lines and lumped elements can be easily obtained by multiplying the corresponding $ABCD$ matrices of the individual sections [1]. For a lossless transmission line, the $ABCD$ parameters are

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}_{\text{lossless line}} = \begin{bmatrix} \cos\theta & jZ_0 \sin\theta \\ (j/Z_0)\sin\theta & \cos\theta \end{bmatrix} \tag{6.59}$$

where $\theta = \beta z'$ is the electrical length of the line segment.

© 2006 by Taylor & Francis Group, LLC

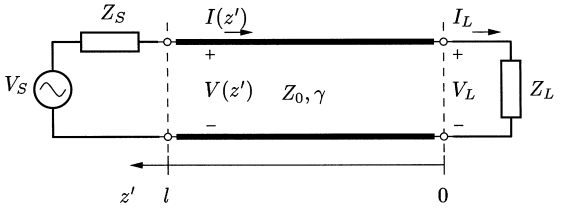


Figure 6.14 Transmission line of finite length terminated in load impedance Z_L.

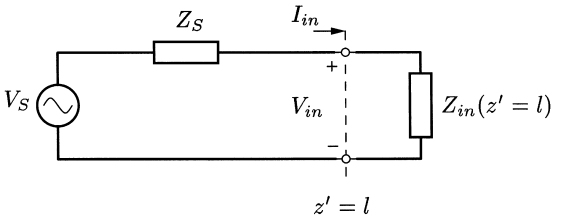


Figure 6.15 Equivalent circuit at the input of the transmission line circuit shown in Fig. 6.14.

determined from the voltage divider circuit. The average power delivered by the source to the input terminals of the transmission line is

$$P_{\text{ave, in}} = \frac{1}{2}\text{Re}\{V_{\text{in}} I_{\text{in}}^*\} \tag{6.61}$$

The average power dissipated in the load impedance $Z_L = R_L + jX_L$ is

$$P_{\text{ave, }L} = \frac{1}{2}\text{Re}\{V_L I_L^*\} = \frac{1}{2}|I_L|^2 R_L = \frac{1}{2}\left|\frac{V_L}{Z_L}\right|^2 R_L \tag{6.62}$$

where V_L and I_L can be determined from the inverse of the $ABCD$ matrix transformation Eq. (6.58).* In general, $P_{\text{ave, }L} < P_{\text{ave, in}}$ for a lossy line and $P_{\text{ave, }L} = P_{\text{ave, in}}$ for a lossless line.

Example. Consider a 10 m long low-loss coaxial cable of nominal characteristic impedance $Z_0 = 75\,\Omega$, attenuation constant $\alpha = 2.2$ dB per 100 ft at 100 MHz, and velocity factor of 78%. The line is terminated in $Z_L = 100\,\Omega$, and the circuit is operated at $f = 100$ MHz. The $ABCD$ parameters for the transmission line are $A = D = -0.1477 + j0.0823$, $B = (-0.9181 + j74.4399)\,\Omega$, and $C = (-0.0002 + j0.0132)\,\Omega^{-1}$. The input impedance of the line is found as $Z_{\text{in}} = (59.3 + j4.24)\,\Omega$. For a source voltage $|V_S| = 10$ V and source impedance $Z_S = 75\,\Omega$, the average power delivered to the input of the line is $P_{\text{ave,in}} = 164.2$ mW and the average power dissipated in the load impedance is $P_{\text{ave, }L} = 138.3$ mW. The difference of 25.9 mW ($\approx 16\%$ of the input power) is dissipated in the transmission line.

*The inverse of Eq. (6.58) expressing the voltage and current at the load in terms of the input voltage and current is

$$\begin{bmatrix} V_L \\ I_L \end{bmatrix} = \begin{bmatrix} D & -B \\ -C & A \end{bmatrix} \begin{bmatrix} V_{\text{in}} \\ I_{\text{in}} \end{bmatrix} \tag{6.63}$$

© 2006 by Taylor & Francis Group, LLC

Transmission Lines as Reactive Circuit Elements

In many practical transmission-line applications, transmission-line losses are small and often negligible. In particular, short sections of transmission lines used as circuit elements in high-frequency circuits are often assumed to be lossless.

For a lossless line with $\gamma = j\beta$ and terminated in a complex load impedance Z_L, the input impedance is

$$Z_{\text{in}}(\theta) = Z_0 \frac{Z_L + jZ_0 \tan\theta}{Z_0 + jZ_L \tan\theta} \tag{6.64}$$

where $\theta = \beta z' = 2\pi z'/\lambda$ is the electrical distance from the termination. Two particularly important special cases are the short-circuited line with $Z_L = 0$ and the open-circuited line with $Z_L \to \infty$.

The input impedance of an open-circuited lossless transmission line is

$$Z_{\text{oc}} = -jZ_0 \cot\theta = jX_{\text{oc}} \tag{6.65}$$

which is purely reactive. The normalized reactance is plotted in Fig. 6.16a. For small line lengths of less than a quarter wavelength ($\theta < 90^\circ$), the input impedance is purely

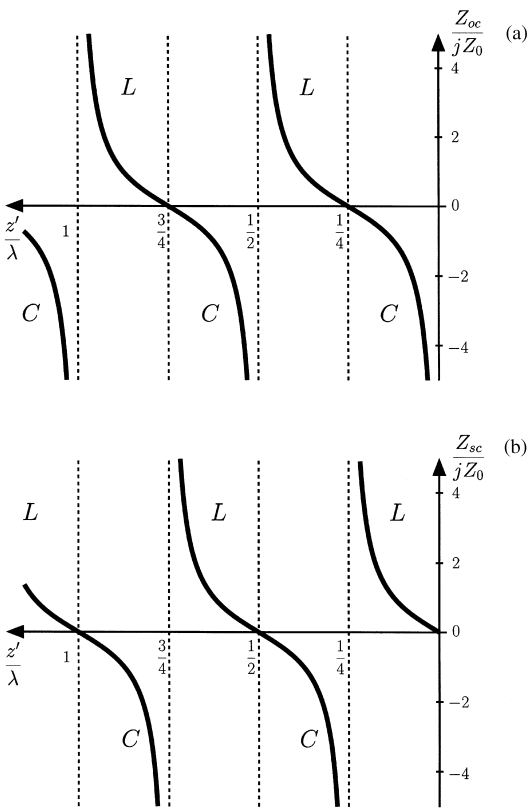


Figure 6.16 Normalized input reactance of a lossless transmission line terminated in (a) an open circuit and (b) a short circuit.

© 2006 by Taylor & Francis Group, LLC

capacitive, as expected. With increasing electrical distance θ, the input reactance alternates every quarter wavelength between being capacitive and inductive. Any reactance value $-\infty < X_{oc} < +\infty$ can be achieved by appropriately adjusting the electrical length (i.e., by varying the physical length or the frequency (wavelength)). Furthermore, for line lengths corresponding to multiples of a half wavelength, the input impedance is again an open circuit. In contrast, for line lengths corresponding to odd multiples of a quarter wavelength, the input impedance is zero [$Z_{oc}(z' = \lambda/4 + n\lambda/2) = 0$, $n = 0, 1, 2, \ldots$].

The input impedance of a short-circuited lossless transmission line is also purely reactive and is given by

$$Z_{sc} = jZ_0 \tan\theta = jX_{sc} \tag{6.66}$$

Figure 6.16b shows the normalized reactance X_{sc}/Z_0 as a function of electrical length θ. For small line lengths of less than a quarter-wavelength ($\theta < 90^\circ$), the input impedance of a short-circuited line is purely inductive, as expected. The dependence of the input reactance of the short-circuited line on electrical length θ corresponds to that of the open-circuited line with a shift by a quarter wavelength. In particular, for line lengths corresponding to multiples of $\lambda/2$, the input impedance is zero, whereas for line lengths corresponding to odd multiples of $\lambda/4$, the input impedance of a short-circuited lossless line is an open circuit [$Z_{sc}(z' = \lambda/4 + n\lambda/2) \to \infty$, $n = 0, 1, 2, \ldots$].

An important application of open- and short-circuited transmission lines is the realization of reactive circuit elements for example for matching networks and filters, in particular at microwave frequencies ranging from a few gigahertz to tens of gigahertz.* At these frequencies, ordinary lumped elements become exceedingly small and difficult to realize and fabricate. In contrast, open- and short-circuited transmission-line sections with lengths on the order of a quarter wavelength become physically small enough to be realized at microwave frequencies and can be easily integrated in planar circuit technology. In practice, it is usually easier to make a good short-circuit termination than an open-circuit termination because of radiation from the open end and coupling to nearby conductors.

Example. To illustrate the design of reactive transmission-line segments, an equivalent inductance $L_{eq} = 5\,\text{nH}$ and an equivalent capacitance $C_{eq} = 2\,\text{pF}$ are realized

*Open- and short-circuit input impedance measurements for a general lossy transmission line can also be used to determine the transmission-line parameters. From $Z_{oc} = Z_0 \coth\gamma z'$ and $Z_{sc} = Z_0 \tanh\gamma z'$ for a lossy line follows

$$Z_0 = \sqrt{Z_{oc} Z_{sc}} \tag{6.67}$$

and

$$\tanh\gamma z' = \sqrt{\frac{Z_{sc}}{Z_{oc}}} \tag{6.68}$$

However, care should be taken in the extraction of $\gamma = \alpha + j\beta$ from Eq. (6.68) due to the periodicity of the phase term $\beta z'$, which must be approximately known.

© 2006 by Taylor & Francis Group, LLC

at $f = 5\,\text{GHz}$ using a short-circuited 50-Ω microstrip line with effective dielectric constant $\epsilon_{\text{eff}} = 1.89$. From Eq. (6.66) follows

$$L_{\text{eq, sc}} \frac{Z_0 \tan\theta_L}{\omega} \tag{6.69}$$

$$C_{\text{eq, sc}} = -\frac{1}{\omega Z_0 \tan\theta_C} \tag{6.70}$$

The minimum electrical lengths for positive values for L_{eq} and C_{eq} are found as $\theta_L = 72.3^\circ$ $(l_L/\lambda = 0.201)$ and $\theta_C = 162.3^\circ$ $(l_C/\lambda = 0.451)$. With $\lambda = 4.36\,\text{cm}$ the corresponding physical lengths of the short-circuited microstrip segments are $l_L = 0.88\,\text{cm}$ and $l_C = 1.97\,\text{cm}$.

Complex Reflection Coefficient

The behavior of a terminated line is further examined in terms of incident and reflected waves at the termination. The ratio of the voltage phasors V^- and V^+ at the termination is defined as the voltage reflection coefficient $\Gamma_L = V^-/V^+$ and is given in terms of the load impedance Z_L and characteristic impedance Z_0 as

$$\Gamma_L = |\Gamma_L| e^{j\theta_L} = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{6.71}$$

The load reflection coefficient Γ_L is in general complex. Here, a different symbol than in Eq. (6.23) is used to emphasize the definition of the complex reflection coefficient as ratio of voltage phasors. For a passive load $|\Gamma_L| \leq 1$. If the terminating load impedance equals the characteristic impedance of the line (matched termination), $\Gamma_L = 0$ and $V^- = 0$. For an open-circuit termination, $\Gamma_L = \Gamma_{\text{oc}} = +1$, while for a short-circuit termination, $\Gamma_L = \Gamma_{\text{sc}} = -1$. In general, for a purely reactive termination $Z_L = jX_L$ ($X_L > 0$ or $X_L < 0$) and real characteristic impedance, the magnitude of the reflection coefficient is $|\Gamma_L| = 1$.

Standing Waves

The total voltage and current along a lossless transmission line with $\gamma = j\beta$ can be expressed with reflection coefficient Γ_L at the termination as

$$V(z') = V_0^+ \{e^{+j\beta z'} + \Gamma_L e^{-j\beta z'}\} \tag{6.72}$$

$$I(z') = \frac{V^+}{Z_0} \{e^{+j\beta z'} - \Gamma_L e^{-j\beta z'}\} \tag{6.73}$$

The superposition of the two opposing traveling wave components leads to periodic variations in voltage and current along the line due to constructive and destructive wave interference. The resulting wave interference component is known as a *standing wave*. For an arbitrary termination with reflection coefficient $\Gamma_L = |\Gamma_L| e^{j\theta_L}$, the voltage and current

© 2006 by Taylor & Francis Group, LLC

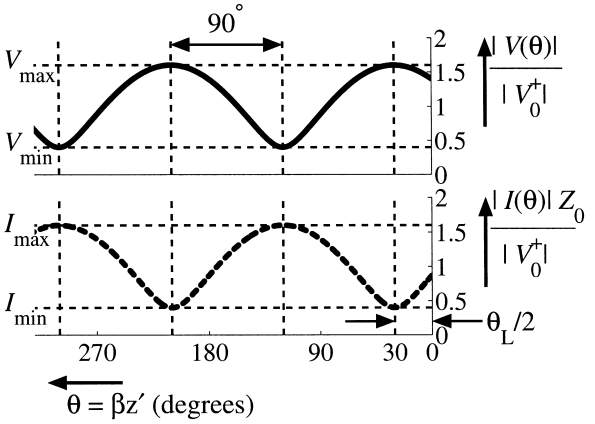


Figure 6.17 Voltage and current standing-wave patterns for a lossless transmission line terminated in a complex load impedance with $\Gamma_L = 0.6\,e^{j60^\circ}$.

standing-wave patterns are given by

$$|V(z')| = |V_0^+|\sqrt{(1+|\Gamma_L|)^2\cos^2\left(\beta z' - \frac{\theta_L}{2}\right) + (1-|\Gamma_L|)^2\sin^2\left(\beta z' - \frac{\theta_L}{2}\right)} \tag{6.74}$$

$$|I(z')| = \frac{|V_0^+|}{Z_0}\sqrt{(1-|\Gamma_L|)^2\cos^2\left(\beta z' - \frac{\theta_L}{2}\right) + (1+|\Gamma_L|)^2\sin^2\left(\beta z' - \frac{\theta_L}{2}\right)} \tag{6.75}$$

Figure 6.17 illustrates the relative voltage and current variations along a lossless transmission line for a general complex load impedance with $\Gamma_L = 0.6e^{j60^\circ}$. In general, the standing-wave pattern on a lossless transmission line is periodic with a period of $\lambda/2$ (or 180° in θ). The voltage magnitude alternates between the maximum and minimum values $V_{\max}$ and $V_{\min}$ given by

$$V_{\max} = (1+|\Gamma_L|)|V_0^+| \tag{6.76}$$

$$V_{\min} = (1-|\Gamma_L|)|V_0^+| \tag{6.77}$$

Similarly, the maximum and minimum current values $I_{\max}$ and $I_{\min}$ are

$$I_{\max} = (1+|\Gamma_L|)\frac{|V_0^+|}{Z_0} = \frac{V_{\max}}{Z_0} \tag{6.78}$$

$$I_{\min} = (1-|\Gamma_L|)\frac{|V_0^+|}{Z_0} = \frac{V_{\min}}{Z_0} \tag{6.79}$$

The locations with maximum voltage can be found from the condition $\beta z' - \theta_L/2 = n\pi$ ($n = 0, 1, 2, \ldots$). The minimum voltages are located a quarter wavelength from the maximum voltages. Locations with current maximum correspond to locations with minimum voltage and vice versa.

© 2006 by Taylor & Francis Group, LLC

The ratio of $V_{\max}$ to $V_{\min}$ is defined as the *standing-wave ratio* (SWR, or S for short) and is given in terms of the reflection coefficient at the termination by

$$\text{SWR} = \frac{V_{\max}}{V_{\min}} = \frac{I_{\max}}{I_{\min}} = \frac{1 + |\Gamma_L|}{1 - |\Gamma_L|} \tag{6.80}$$

The standing-wave ratio is a measure for the amount of mismatch at the termination. The standing-wave ratio for a matched termination is SWR = 1. For an open-circuit, a short-circuit, or a purely reactive termination SWR $\to \infty$. For a resistive or complex termination, $1 < \text{SWR} < \infty$. In general, SWR varies in the range

$$1 \leq \text{SWR} \leq \infty \tag{6.81}$$

Table 6.3 shows the standing-wave patterns for several special types of terminations. For an open-circuit termination and a purely resistive termination with $R_L > Z_0$, the voltage is maximum at the termination. In contrast, the voltage at the termination is minimum for a short-circuit termination or a purely resistive termination with $R_L < Z_0$. A resistive termination causes a compression in the standing-wave pattern, whereas a reactive termination gives rise to a shift of the voltage maximum away from the termination. For a complex termination as shown in Fig. 6.17 with $\Gamma_L = 0.6e^{j60^\circ}$, the standing-wave pattern is both compressed (SWR = 4) and shifted toward the source side by $\theta_L/2 = +30^\circ$ compared to the open-circuit case.

The standing-wave ratio and the distance from the termination to the nearest voltage maximum can be determined in an experimental setup to find the complex reflection coefficient and, hence, the complex impedance of an unknown termination.* The reflection coefficient magnitude $|\Gamma_L|$ is given in terms of SWR as

$$|\Gamma_L| = \frac{\text{SWR} - 1}{\text{SWR} + 1} \tag{6.82}$$

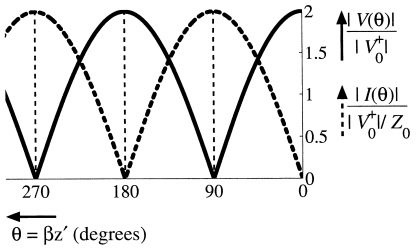
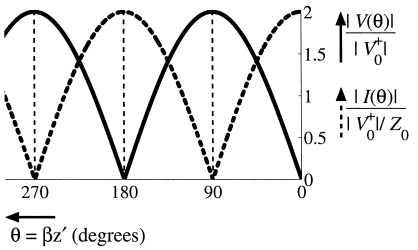
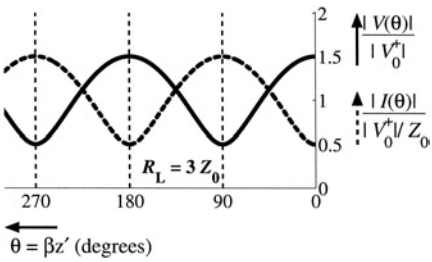
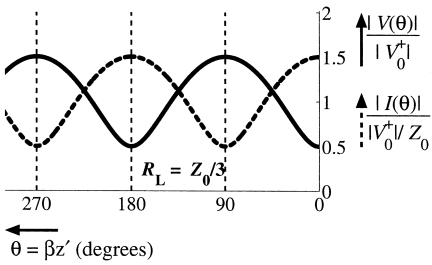
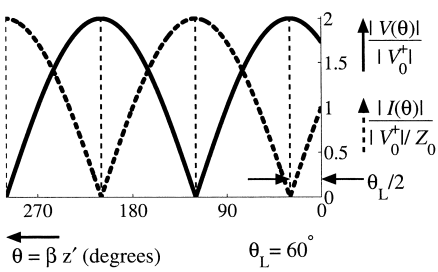
Example. From standing-wave measurements, the standing-wave ratio is found as $\text{SWR} = V_{\max}/V_{\min} = 5$, the distance between successive voltage minima is 20 cm, and the distance from the termination to the nearest voltage minimum is 4 cm. From Eq. (6.82) follows the magnitude of the reflection coefficient at the termination as $|\Gamma_L| = (5-1)/(5+1) = 2/3$. The wavelength on the line corresponds to twice the distance between successive voltage minima and is $\lambda = 40$ cm. The distance from the termination to the closest voltage minimum is $4/40\,\lambda = \lambda/10$ or 36°, and the distance to the nearest voltage maximum is $\lambda/10 + \lambda/4 = 0.35\,\lambda$ or 126°. The phase of the reflection coefficient is $\theta_L = 2 \times 126^\circ = 252^\circ$. The corresponding load impedance is found with $Z_L = Z_0(1 + \Gamma_L)/(1 - \Gamma_L)$ as $Z_L = (0.299 - j0.683)Z_0$.

In most applications, the phase information for the reflection coefficient is not needed. The magnitude of the reflection coefficient directly determines the fraction of

*In practice, it is easier to accurately determine the location of a voltage minimum. The location to the voltage maximum can be obtained from the location of the voltage minimum by adding or subtracting a quarter wavelength.

© 2006 by Taylor & Francis Group, LLC

Table 6.3 Standing-wave Patterns on a Lossless Transmission Line for Special Types of Terminations

Type of termination	Standing-wave pattern
Open circuit $\lvert V(z')\rvert = 2\lvert V_0^+\rvert\lvert\cos\beta z'\rvert$ $\lvert I(z')\rvert = \dfrac{\lvert V_0^+\rvert}{Z_0}\lvert\sin\beta z'\rvert$ $\Gamma_L = +1 \quad \text{SWR} = \infty$	
Short circuit $\lvert V(z')\rvert = 2\lvert V_0^+\rvert\lvert\sin\beta z'\rvert$ $\lvert I(z')\rvert = 2\dfrac{V_0^+}{Z_0}\lvert\cos\beta z'\rvert$ $\Gamma_L = -1 \quad \text{SWR} = \infty$	
Resistive termination $R_L > Z_0$ $\lvert V(z')\rvert = \dfrac{2}{R_L + Z_0}\lvert V_0^+\rvert\sqrt{R_L^2\cos^2\beta z' + Z_0^2\sin^2\beta z'}$ $\lvert I(z')\rvert = \dfrac{2}{R_L + Z_0}\dfrac{\lvert V_0^+\rvert}{Z_0}\sqrt{Z_0^2\cos^2\beta z' + R_L^2\sin^2\beta z'}$ $\Gamma_L = \dfrac{R_L - Z_0}{R_L + Z_0} > 0 \quad \text{SWR} = \dfrac{R_L}{Z_0}$	
Resistive termination $R_L < Z_0$ $\lvert V(z')\rvert = \dfrac{2}{R_L + Z_0}\lvert V_0^+\rvert\sqrt{Z_0^2\cos^2\beta z' + R_L^2\sin^2\beta z'}$ $\lvert I(z')\rvert = \dfrac{2}{R_L + Z_0}\dfrac{\lvert V_0^+\rvert}{Z_0}\sqrt{R_L^2\cos^2\beta z' + Z_0^2\sin^2\beta z'}$ $\Gamma_L = \dfrac{R_L - Z_0}{R_L + Z_0} < 0 \quad \text{SWR} = \dfrac{Z_0}{R_L}$	
Reactive termination $Z_L = jX_L$ $\lvert V(z')\rvert = 2\lvert V_0^+\rvert\left\lvert\cos(\beta z' - \theta_L/2)\right\rvert$ $\lvert I(z')\rvert = 2\dfrac{\lvert V_0^+\rvert}{Z_0}\left\lvert\sin(\beta z' - \theta_L/2)\right\rvert$ $\Gamma_L = 1e^{j\theta_L}$ $\theta_L = 2\tan^{-1}(Z_0/X_L) \quad \text{SWR} \to \infty$	

© 2006 by Taylor & Francis Group, LLC

average incident power that is reflected back on the transmission line. With Eqs. (6.72) and (6.73), the net power flow on a lossless transmission line is given by

$$P_{\text{ave}}(z') = \frac{1}{2}\text{Re}\{V(z')I^*(z')\} = \frac{|V_0^+|^2}{2Z_0}\left(1 - |\Gamma_L|^2\right) = P_{\text{ave}}^+\left(1 - |\Gamma_L|^2\right) \tag{6.83}$$

which is independent of position z' on the line. The fraction of average incident power P_{ave}^+ that is reflected is

$$P_{\text{ave}}^- = -|\Gamma_L|^2 P_{\text{ave}}^+ \tag{6.84}$$

The negative sign in Eq. (6.84) indicates the power flow away from the load. Note that the incident power P_{ave}^+ is the combined power due to all forward traveling wave components and thus depends on the load impedance if the source is not matched ($Z_S \neq Z_0$).

In many transmission systems, such as a radio transmitter site, it is critical to monitor the amount of reflected power. The percentage of reflected power can be directly determined from the measured standing-wave ratio. For example, for SWR = 1.5, the magnitude of the reflection coefficient is 0.2, which means that 4% of the incident power is reflected. For a 60-KW transmitter station this would amount to a reflected power of 2400 W.

6.4.3. The Smith Chart

The *Smith chart*, developed by P. H. Smith in 1939, is a powerful graphical tool for solving and visualizing transmission-line problems [17,18]. Originally intended as a graphical transmission-line calculator before the computer age to perform calculations involving complex impedances, the Smith chart has become one of the primary graphical display formats in microwave computer-aided design software and in some commonly used laboratory test equipment, in particular the network analyzer.

The transformation of complex impedance along a transmission line given in Eq. (6.64) is mathematically complicated and lacks visualization and intuition. On the other hand, the reflection coefficient undergoes a simple and intuitive transformation along the transmission line. The reflection coefficient at distance z' from the termination is defined as $\Gamma(z') = V^-(z')/V^+(z')$ and is given in terms of the reflection coefficient at the termination Γ_L by

$$\Gamma(z') = \Gamma_L e^{-j2\beta z'} = |\Gamma_L| e^{j(\theta_L - 2\beta z')} \tag{6.85}$$

The magnitude of the reflection coefficient is unchanged along the lossless line and the phase of the reflection coefficient is reduced by twice the electrical distance from the termination.

The Smith chart combines the simple transformation property of the reflection coefficient along the line with a graphical representation of the mapping of normalized

© 2006 by Taylor & Francis Group, LLC

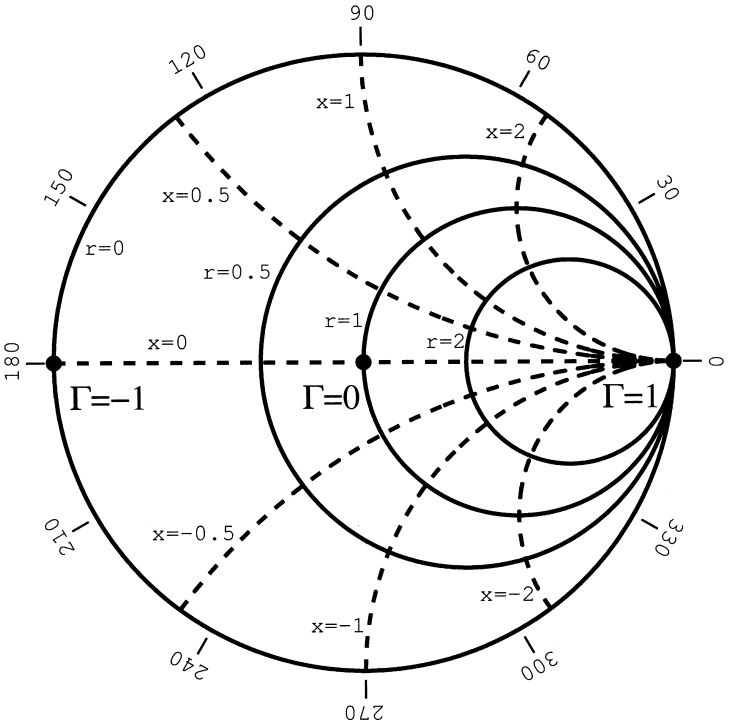


Figure 6.18 Illustration of the basic features of the Smith chart.

impedance to the complex reflection coefficient plane given by

$$\mathsf{z}(z') = \frac{Z_{\text{in}}(z')}{Z_0} = \frac{1+\Gamma(z')}{1-\Gamma(z')} \tag{6.86}$$

Here, $\mathsf{z} = \mathsf{r} + j\mathsf{x} = Z/Z_0$ is defined as the normalized impedance with respect to the characteristic impedance of the line. The combination of these two operations in the Smith chart enables the simple graphical determination and visualization of the impedance transformation along a transmission line. Other parameters, such as the standing-wave ratio or the locations of voltage maxima and minima on the line can be simply read off the Smith chart, and more advanced transmission-line calculations and circuit designs can be performed with the Smith chart.

Figure 6.18 illustrates the basic features of the Smith chart. The chart shows a grid of normalized impedance coordinates plotted in the complex plane of the reflection coefficient. The impedance grid consists of a set of circles for constant values of normalized resistance r and a set of circular arcs for constant values of normalized reactance x. Any normalized impedance $\mathsf{z} = \mathsf{r} + j\mathsf{x}$ on a transmission line corresponds to a particular point on or within the unit circle ($|\Gamma| = 1$ circle) in the complex plane of the reflection coefficient. For a matched impedance the $\mathsf{r} = 1$ circle and $\mathsf{x} = 0$ line intersect at the origin of the Smith chart ($\Gamma = 0$). The open-circuit point $\Gamma = 1$ is to the far right, and the short-circuit point $\Gamma = -1$ is to the far left, as indicated in Fig. 6.18. In a real Smith chart, as shown

© 2006 by Taylor & Francis Group, LLC

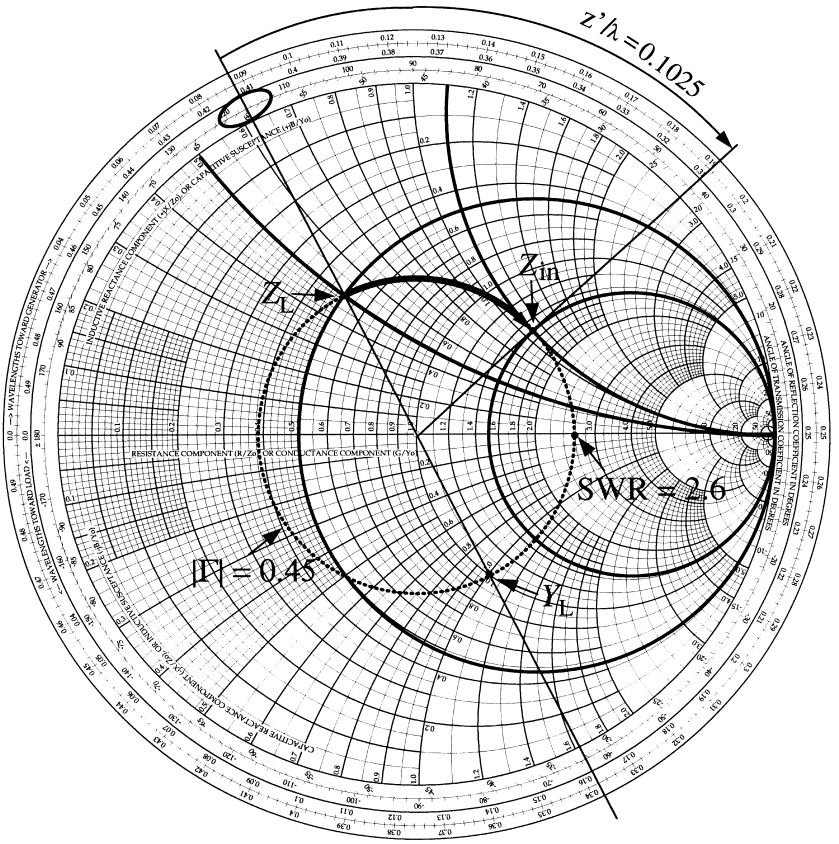


Figure 6.19 Smith chart example for $Z_L = (25 + j25)\,\Omega$ and $Z_0 = 50\,\Omega$.

in Fig. 6.19, a fine grid is used for added accuracy, and scales are added to help with the calculation of phase change in reflection coefficient along the transmission line.

Example. To illustrate the use of the Smith chart for transmission-line calculations, consider a lossless line with characteristic impedance $Z_0 = 50\,\Omega$, which is terminated in a complex load impedance $Z_L = (25 + j25)\,\Omega$. The normalized load impedance is $\mathsf{z} = 0.5 + j0.5$ and is shown on the Smith chart as the intersection of the $\mathsf{r} = 0.5$ and $\mathsf{x} = 0.5$ grid circles. The load reflection coefficient can be directly read off from the Smith chart. The radius of the transformation circle through Z_L (relative to the radius of the unit circle $\mathsf{r} = 0$) gives the magnitude of the reflection coefficient as $|\Gamma_L| = 0.45$. The phase of the reflection coefficient is $\theta_L = 116.5^\circ$. The standing-wave ratio on the line corresponds to the normalized maximum impedance $\mathsf{z}_{\max}$ along the line, which is real and lies on the intersection of the transformation circle and the $\mathsf{x} = 0$ line. The standing-wave ratio can be directly read off the Smith chart as SWR $= 2.6$. For a given electrical length of the line, the input impedance is found by first determining the reflection coefficient at the input through clockwise rotation on the transformation (SWR) circle by twice the electrical length of the line, as given by Eq. (6.85). Assuming an electrical length of $l = 0.1025\,\lambda$, the phase of the reflection coefficient changes by $-2\beta l = -4\pi \times 0.1025$. This amounts to a rotation in *clockwise* direction by about 74°. For convenience, the Smith chart includes scales around its periphery, which can be used to determine the amount of phase rotation directly in

© 2006 by Taylor & Francis Group, LLC

units of wavelengths. In this example, the starting value at the load on the rotation scale labeled "toward generator" is 0.088. The end value is $0.088 + 0.1025 = 0.1905$. The phase of the reflection coefficient is read off as $\theta_{\text{in}} = 42.5^\circ$. Finally, the input impedance is obtained as the intersection of the line through the origin with constant phase and the transformation circle. The normalized input impedance is approximately found as $\mathsf{z}_{\text{in}} = 1.5 + j1.1$, or $Z_{\text{in}}(z' = l = 0.1025\lambda) = (75 + j55)\,\Omega$.

In transmission-line problems with parallel-connected elements, it is advantageous to work with admittances rather than impedances. The impedance Smith chart can be conveniently used with normalized admittances $\mathsf{y} = \mathsf{g} + j\mathsf{b} = YZ_0$ by considering the relationship

$$\Gamma = \frac{\mathsf{z} - 1}{\mathsf{z} + 1} = -\frac{\mathsf{y} - 1}{\mathsf{y} + 1} \tag{6.87}$$

where $\mathsf{y} = 1/\mathsf{z}$. This relationship shows that the impedance grid can be directly used as admittance grid with $\mathsf{g} = \text{const}$ circles and $\mathsf{b} = \text{const}$ circular arcs if the reflection coefficient is multiplied by negative one, which amounts to a rotation by 180° on the Smith chart. Then, the open circuit is located at the far left and the short circuit is at the far right. The conversion from normalized impedance coordinates to normalized admittance coordinates given by $\mathsf{y} = 1/\mathsf{z}$ can be simply achieved on the Smith chart by a 180° rotation along the transformation (SWR) circle. For example, for the normalized load impedance $\mathsf{z} = 0.5 + j0.5$, the normalized load admittance is found as $\mathsf{y} = 1/\mathsf{z} = 1 - j$, as indicated in Fig. 6.19.

6.4.4. Impedance Matching

In many transmission-line applications, it is desirable to match the load impedance to the characteristic impedance of the line and eliminate reflections in order to maximize the power delivered to the load and minimize signal distortion and noise.* Reducing or eliminating reflections from the load is particularly important in high-power RF transmission systems to also minimize hot spots along a transmission line (e.g., the feed line between the transmitter and the antenna) that are caused by standing waves and not exceed the power-handling capabilities of the transmission line. Excessive reflections can also damage the generator, especially in high-power applications.

In practice, the impedance of a given load is often different from the characteristic impedance Z_0 of the transmission line, and an additional impedance transformation network is needed to achieve a matched load condition. Figure 6.20 illustrates the basic idea of matching an arbitrary load impedance Z_L to a transmission line. The matching network is designed to provide an input impedance looking into the network that is equal to Z_0 of the transmission line, and thus eliminate reflections at the junction between the transmission line and the matching network. The matching network is ideally lossless so

*In general, impedance matching can be done at the load or the source end, or at both ends of the transmission line. For a matched source, maximum power is delivered to the load when it is matched to the transmission line and power loss on the line is minimized. For a given source impedance Z_S, maximum power transmission on a lossless line is achieved with conjugate matching at the source ($Z_{\text{in}} = Z_S^*$) [1].

© 2006 by Taylor & Francis Group, LLC

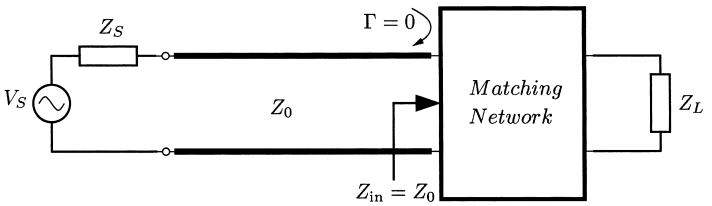


Figure 6.20 General illustration of impedance matching at the termination.

that all incident power on the line ends up being dissipated in the load. A lossless matching network may consist of lumped reactive elements or reactive transmission-line elements (stubs) at higher frequencies and/or cascaded transmission-line sections of appropriate length.

A matching network requires at least two adjustable parameters, such as a lumped series element and a lumped shunt element, each with adjustable reactance value, to independently transform the real and imaginary parts of the load impedance. Because the elements in the matching network are frequency dependent, the exact matching condition is generally achieved only at a single design frequency. For other frequencies, the reflection coefficient will be sufficiently small only over a narrow bandwidth about the design frequency. Larger matching bandwidths may be achieved if more independent elements are used in the matching network.

Many different design choices of matching networks are available. The selection of a particular matching network may depend on a number of factors including realizability in a given technology, required bandwidth, simplicity, occupied space, tunability of the matching network, and cost of implementation. In the following, two common matching methods using sections of lossless transmission lines are described to further illustrate the concept of impedance matching.

Quarter-wave Transformer

A lossless transmission line of length $l = \lambda/4$ has a special simplified impedance transformation property, which can be advantageously used for impedance matching. With Eq. (6.64), the input impedance of a lossless transmission line of length $l = \lambda/4$ and characteristic impedance $Z_{0,\mathrm{T}}$ that is terminated with load impedance Z_L is

$$Z_{\text{in}}|_{l=\lambda/4} = \frac{Z_{0,\mathrm{T}}^2}{Z_L} \tag{6.88}$$

In particular, any purely resistive load impedance $Z_L = R_L$ is transformed into a resistive input impedance given by $R_{\text{in}} = Z_{0,\mathrm{T}}^2/R_L$. Hence, a quarter-wave section of a transmission line can be directly used to match a purely resistive load impedance R_L to a line with characteristic impedance Z_0 if the characteristic impedance $Z_{0,\mathrm{T}}$ of the quarter-wave section is given by

$$Z_{0,\mathrm{T}} = \sqrt{R_L Z_0} \tag{6.89}$$

For example, to match a half-wave dipole antenna with input impedance $Z_L \approx 73\,\Omega$ to a twin-lead cable with $Z_0 = 300\,\Omega$, the characteristic impedance of the quarter-wave transformer should be $Z_{0,\mathrm{T}} = \sqrt{73\,\Omega \cdot 300\,\Omega} \approx 148\,\Omega$.

© 2006 by Taylor & Francis Group, LLC

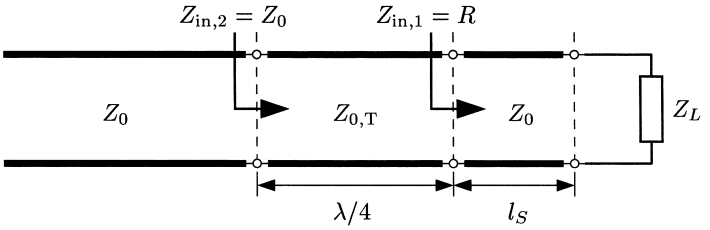


Figure 6.21 Impedance matching of a complex load using a quarter-wave transformer.

If the load impedance is complex, it is necessary to first transform the complex impedance to a real impedance. This can be accomplished with a section of transmission line of appropriate length l_s between the load and the quarter-wave transform, as illustrated in Fig. 6.21. A transmission line can transform any complex load impedance with $|\Gamma_L| < 1$ to a resistive impedance at the locations with either voltage maximum or voltage minimum. The transmission-line transformation of a complex load to a real impedance is best illustrated on the Smith chart. For example, consider a complex load consisting of a parallel combination of $R_L = 125\,\Omega$ and $C_L = 2.54\,\text{pF}$. At the design frequency $f_0 = 1\,\text{GHz}$, the load impedance is $Z_L = (25 - j50)\,\Omega$. The normalized load impedance $\mathsf{z}_L = 0.5 - j$ for $Z_0 = 50\,\Omega$ is shown on the Smith chart in Fig. 6.22. The transformation circle intersects the $\mathsf{x} = 0$ grid line at $r_{\min} \approx 0.24$ and $r_{\max} = 1/r_{\min} \approx 4.2$. The distance to the closest location with real input impedance ($\mathsf{z}_{\text{in}} = r_{\min}$) is found as $l_s = 0.135\,\lambda$. The input impedance at this location is $Z_{\text{in},1} = R = r_{\min} Z_0 \approx 12\,\Omega$, and the characteristic impedance of the quarter-wave transformer is found as $Z_{0,\text{T}} = \sqrt{RZ_0} \approx 24.5\,\Omega$. The second solution with real input impedance is at the voltage maximum with $R = r_{\max} Z_0 \approx 210\,\Omega$ and $l_s = 0.135\,\lambda + 0.25\,\lambda = 0.385\,\lambda$. The corresponding characteristic impedance of the quarter-wave transformer is $Z_{0,\text{T}} = \sqrt{RZ_0} \approx 102.5\,\Omega$. Typically, the solution with the shortest line length l_s is chosen unless it is difficult to realize the characteristic impedance of the corresponding quarter-wave transformer.

Figure 6.23 shows the response of the matching network as a function of frequency. The matching network gives an exact match (SWR = 1) at the design frequency $f_0 = 1\,\text{GHz}$. The bandwidth defined here as the frequency band around the center frequency with SWR ≤ 1.5 is about 100 MHz or 10%. The standing-wave ratio response without matching network is also shown in Fig. 6.23 for comparison.

The bandwidth of the matching network can be increased, for example, by cascading multiple quarter-wave sections (multisection quarter-wave transformer) with smaller impedance steps per section giving an overall more gradual impedance transformation [1]. This type of matching network can be easily implemented in planar transmission line technology, such as microstrip, where the characteristic impedance can be changed continuously by varying the line width or spacing.

Stub Matching

In another common impedance matching technique, a reactive element of appropriate value is connected either in series or in parallel to the transmission line at a specific distance from the load. The reactive element can be realized as open- or short-circuited

© 2006 by Taylor & Francis Group, LLC

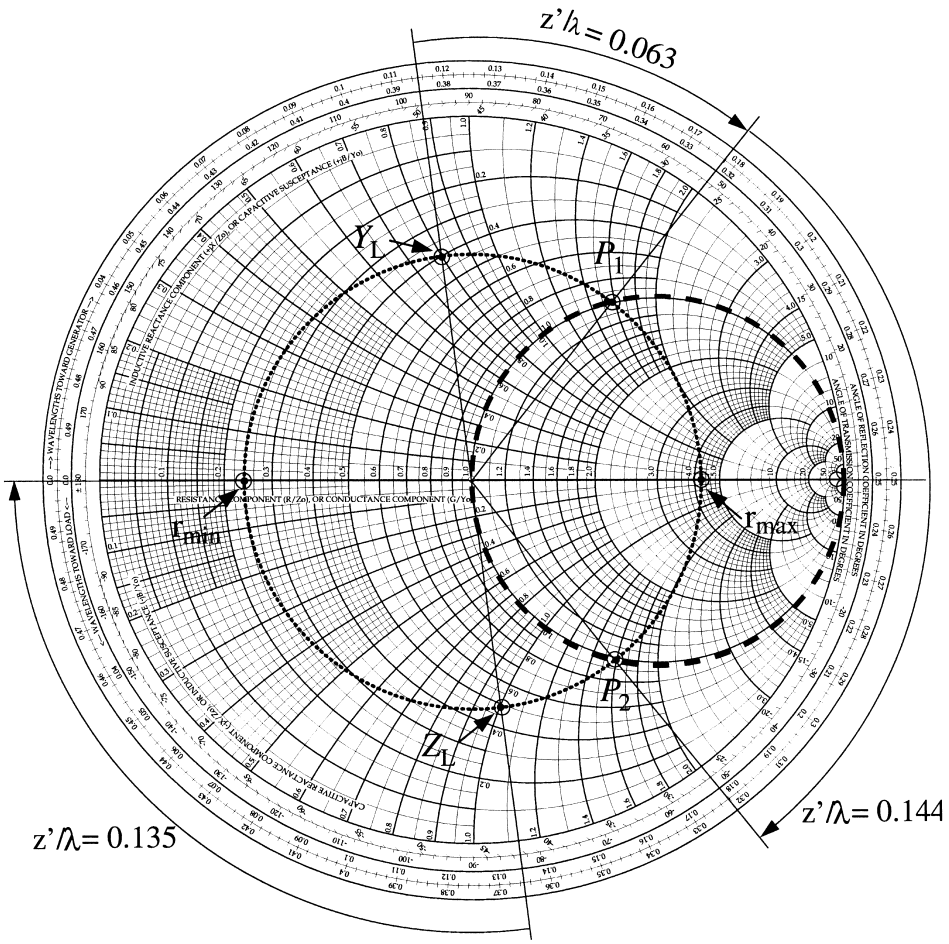


Figure 6.22 Graphical illustration on the Smith chart of quarter-wave matching and shunt (stub) matching of a complex load impedance $Z_L/Z_0 = 0.5 - j$.

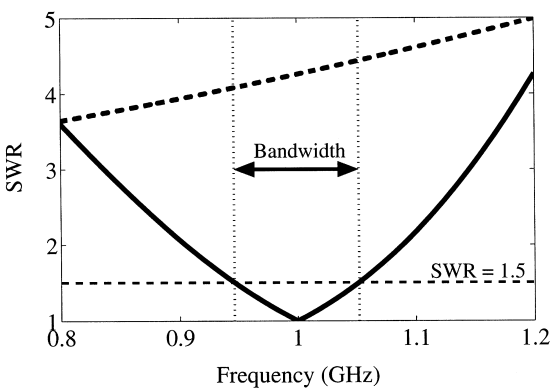


Figure 6.23 SWR = 1.5 bandwidth of an example matching network using a quarter-wave transformer. Also shown with a dashed line is the response without the matching network.

© 2006 by Taylor & Francis Group, LLC

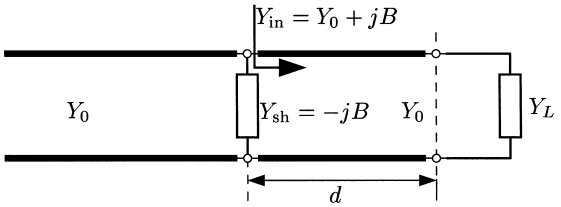


Figure 6.24 Matching network with a parallel shunt element.

stub or as lumped inductor or capacitor element. The two design parameters of a stub matching network are the distance from the termination at which the reactive element is connected, and the stub length needed to realize the required reactance.

The general matching procedure with a single reactive element or a stub is demonstrated for a parallel (shunt) configuration with shunt admittance element Y_{sh} connected at distance d from the termination, as illustrated in Fig. 6.24. For shunt connections it is more convenient to work with admittances than with impedances. The transmission line transforms the load admittance $Y_L = 1/Z_L$ to an input admittance $Y_{\text{in}} = G + jB$ at distance d from the termination. In the first step of the matching procedure, distance d is selected such that the real part of the input admittance is matched as $G = Y_0$, and the nonzero input susceptance B is determined. In the second step, a reactive shunt element with admittance $Y_{\text{sh}} = -jB$ is added to cancel out susceptance B in the input admittance. The summation of shunt admittance and input admittance of the line yields a matched total admittance $Y_0 = 1/Z_0$.

The shunt matching procedure is further illustrated on the Smith chart shown in Fig. 6.22. The same normalized load impedance $\mathsf{z}_L = 0.5 - j$ as in the previous matching network example is assumed. The corresponding normalized load admittance is found from the Smith chart as $\mathsf{y}_L = 0.4 + j0.8$. The transformation circle with $|\Gamma| = \text{const}$ intersects the $\mathsf{g} = 1$ circle at two points labeled as P_1 and P_2 satisfying the condition $\mathsf{y}_{\text{in}} = 1 + jb$. Any complex load admittance with $|\Gamma_L| < 1$ can be transformed by a transmission line of appropriate length to a point on the $\mathsf{g} = 1$ circle. The normalized input admittances at points P_1 and P_2 are $\mathsf{y}_{\text{in},1} = 1 + j1.58$ and $\mathsf{y}_{\text{in},2} = 1 - j1.58$, respectively. The distance from the termination to point P_1 on the line with matched real part of the input admittance is found as $d_1 = 0.063\,\lambda$. The distance to P_2 is $d_2 = d_1 + 0.144\,\lambda = 0.207\,\lambda$. The normalized input susceptance $b_1 = 1.58$ at position P_1 is capacitive and needs to be canceled with an inductive shunt element with normalized admittance $\mathsf{y}_{\text{sh}} = -j1.58$. The required shunt element may be realized with a lumped inductor or an open- or short-circuited stub of appropriate length. Similarly, matching position P_2 with $\mathsf{y}_{\text{in},2} = 1 - j1.58$ requires a capacitive shunt element to cancel the susceptance. The capacitive shunt admittance may be realized with a lumped capacitor or an open- or short-circuited stub of appropriate length.

6.5. FURTHER TOPICS OF TECHNOLOGICAL IMPORTANCE AND FUTURE DIRECTIONS

In this section, further transmission-line topics of technological importance are briefly discussed and current developments and future directions are outlined.

© 2006 by Taylor & Francis Group, LLC

6.5.1. Coupled Lines

Transmission-line circuits often consist of multiple parallel conductors that are in close proximity to each other. Examples of multiconductor transmission-line systems include multiphase power lines, telephone cables, and data bus lines on the printed-circuit board (PCB) of a digital system. Due to the proximity of the conductors, the time-varying electromagnetic fields generated by the different transmission lines interact, and the lines become capacitively and inductively coupled. The propagation characteristics of coupled lines depend not only on the line parameters of the individual lines but also on the mutual distributed capacitance and inductance parameters.

The capacitive and inductive coupling between transmission lines often leads to adverse effects in a transmission system. As an example, coupling between closely spaced lines (interconnects) in digital systems can lead to unwanted crosstalk noise and generally sets an upper limit in interconnection density (see e.g. Refs. 10 and 19). On the other hand, electromagnetic coupling between adjacent lines can be used to advantage to realize a variety of components for microwave circuits such as filters, directional couplers, and power dividers [1]. Recently, there has also been increased interest in the realization of compact three-dimensional embedded passive components for RF and mixed-signal modules, and new compact designs using coupled lines have been demonstrated (e.g., Ref. 20). A general overview of coupled transmission-line theory and its application to cross-talk analysis and design of passive microwave components is given, e.g., in Ref. 5.

6.5.2. Differential Lines

A differential line can be considered as a special case of two symmetric coupled lines. A differential line consists of two closely spaced symmetric signal conductors that are driven with identical signals of opposite polarity with respect to a common ground reference (differential signaling). The main advantages of differential lines include an increased immunity to common-mode noise and the localized ground references at the input and output of the line. In particular, the net return current in the ground conductor of a differential transmission line is ideally zero, which helps to eliminate or reduce the effects of nonideal current return paths with finite resistance and inductance. As a disadvantage, differential lines require more conductor traces and generally need to be carefully routed to avoid conversion between differential- and common-mode signals. Because of the advantageous properties of differential lines compared to regular (single-ended) lines, however, differential lines are increasingly being used for critical signal paths in high-speed analog and digital circuits (see, e.g., Refs. 10 and 19). Differential circuit architectures are also being employed in parts of RF circuits because of their superior noise-rejection properties [21].

6.5.3. Chip- and Package-level Interconnects

Transmission lines or electrical interconnects are present at various levels of an electronic system ranging from cabling to printed-circuit board level to chip packaging to chip level. The electrical interconnections in an electronic package constitute the electrical interface between the chip (or a set of chips packaged in a module) and the rest of the electronic system. The package interconnections can generally be represented by a combination of

© 2006 by Taylor & Francis Group, LLC

lumped R, L, C elements and nonuniform coupled transmission lines. In some advanced high-performance packages the interconnections are realized in form of a miniature printed wiring board with several levels of metalization. The electronic package may significantly influence the electrical performance of an integrated circuit; hence, the package characteristics should be included in the design phase of the integrated circuit. The co-design of the integrated circuit and package has recently been pursued for both digital and RF integrated circuits as well as for system-on-a-chip solutions.

At the chip level, interconnects in VLSI and RF integrated circuits usually behave as lumped or distributed RC circuits because of the large series resistance of the metalization. With increasing clock frequencies, however, the distributed series inductance becomes more and more significant. As a result, inductance effects cannot be neglected in some of the longer on-chip interconnects in present-day high-performance VLSI circuits [22]. On-chip interconnects with nonnegligible inductance exhibit transmission-line behavior and need to be modeled as lossy transmission lines rather than RC lines.

6.5.4. CAD Modeling of Transmission Lines

The development of dispersive single and coupled transmission-line models for computer-aided design (CAD) tools is an active area of research in both industry and academia. In general, the line parameters of a transmission line are frequency dependent because of conductor loss (including skin and proximity effects), substrate loss, and dispersion induced by inhomogeneous dielectric substrates. The frequency-dependent transmission-line parameters, however, cannot be represented directly in a time-domain simulator environment such as SPICE. Several approaches for modeling lossy dispersive transmission lines have been developed including (1) convolution with the impulse response of the lossy transmission line, (2) synthesis of the frequency-dependent line parameter in terms of ideal lumped elements and controlled sources for a short line section, and (3) mathematical macromodels obtained with model-order reduction (MOR) techniques resulting in an approximation of the transmission-line characteristics with a finite number of pole-residue pairs. Other areas of current and future interest include the efficient extraction of the line parameters (or parasitics) and the cosimulation of the electromagnetic, thermal, and mechanical phenomena in an electronic system. A review of the methodologies for the electrical modeling of interconnects and electronic packages is given in Ref. 23. Modeling of coupled transmission lines–interconnects based on model-order reduction is further described in Ref. 24.

REFERENCES

1. Pozar, D.M. *Microwave Engineering*, 2nd Ed.; Wiley: New York, 1998.
2. Collin, R.E. *Field Theory of Guided Waves*, 2nd Ed.; IEEE Press: New York, 1991.
3. Cheng, D.K. *Field and Wave Electromagnetics*, 2nd Ed.; Addison-Wesley: Reading, MA, 1990.
4. Ramo, S.; Whinnery, J.R; Van Duzer, T. *Fields and Waves in Communication Electronics*, 3rd Ed.; Wiley: New York, 1993.
5. Magnusson, P.C.; Alexander, G.C.; Tripathi, V.K.; Weisshaar, A. *Transmission Lines and Wave Propagation*, 4th Ed.; CRC Press: Boca Raton, FL, 2001.
6. Hoffmann, R.K. *Handbook of Microwave Integrated Circuits*; Artech House: Norwood, MA, 1987.
7. Wadell, B.C. *Transmission Line Design Handbook*; Artech House: Norwood, MA, 1991.

© 2006 by Taylor & Francis Group, LLC

8. Nagel, L.W. SPICE: A computer program to simulate semiconductor circuits, Tech. Rep. ERL-M520, Univ. California, Berkeley, May 1975.
9. Johnson, H.W.; Graham, M. *High-Speed Digital Design: A Handbook of Black Magic*; Prentice-Hall: Englewood Cliffs, NJ, 1993.
10. Hall, S.H.; Hall, G.W.; McCall, J.A. *High-Speed Digital System Design: A Handbook of Interconnect Theory and Design Practices*; Wiley: New York, 2000.
11. Freeman, J.C. *Fundamentals of Microwave Transmission Lines*; Wiley: New York, 1996.
12. DeFalco, J.A. Reflection and cross talk in logic circuit interconnections. IEEE Spectrum **July 1970**, 44–50.
13. Oliver, B.M. Time-domain reflectometry. Hewlett-Packard **Feb**. **1964**, *15* (6), 1–7.
14. Inan, U.S.; Inan, A.S. *Engineering Electromagnetics*; Addison-Wesley: Reading, MA, 1998.
15. Jong, J.M.; Tripathi, V.K. Equivalent circuit modeling of interconnects from time domain measurements. IEEE Trans. Comp., Pack., Manufact. Technol. **Feb**. **1993**, *16* (1), 119–126.
16. Lathi, B.P. *Linear Systems and Signals*; Oxford University Press: New York, 2002.
17. Smith, P.H. Transmission line calculator. Electronics **Jan**. **1939**, *12* (1), 29–31.
18. Smith, P.H. An improved transmission-line calculator. Electronics **Jan**. **1944**, *17* (1), 130, 318.
19. Dally, W.J.; Poulton, J.W. *Digital Systems Engineering*; Cambridge University Press: New York, 1998.
20. Settaluri, R.K.; Weisshaar, A; Tripathi, V.K. Design of compact multilevel folded-line bandpass filters. IEEE Trans. Microwave Theory Tech. **Oct. 2001**, *49* (10), 1804–1809.
21. Lee, T.H. *The Design of CMOS Radio-Frequency Integrated Circuits*; Cambridge University Press: New York, 1998.
22. Deutsch, A. When are transmission-line effects important for on-chip interconnections? IEEE Trans. Microwave Theory Tech. **Oct**. **1997**, *45* (10), 1836–1846.
23. Ruehli, A.E.; Cangellaris, A.C. Progress in the methodologies for the electrical modeling of interconnects and electronic packages. Proc. IEEE **May 2001**, *89* (5), 740–771.
24. Achar, R.; Nakhla, M. Simulation of high-speed interconnects. Proc. IEEE **May 2001**, *89* (5), 693–728.

© 2006 by Taylor & Francis Group, LLC

7
Waveguides and Resonators

Kenneth R. Demarest
The University of Kansas
Lawrence, Kansas

7.1. INTRODUCTION

Any structure that transports electromagnetic waves can be considered as a *waveguide*. Most often, however, this term refers to either metal or dielectric structures that transport electromagnetic energy without the presence of a complete circuit path. Waveguides that consist of conductors and dielectrics (including air or vacuum) are called *metal waveguides*. Waveguides that consist of only dielectric materials are called *dielectric waveguides*.

Metal waveguides use the reflective properties of conductors to contain and direct electromagnetic waves. In most cases, they consist of a long metal cylinder filled with a homogeneous dielectric. More complicated waveguides can also contain multiple dielectrics and conductors. The conducting cylinders usually have rectangular or circular cross sections, but other shapes can also be used for specialized applications. Metal waveguides provide relatively low loss transport over a wide range of frequencies—from RF through millimeter wave frequencies.

Dielectric waveguides guide electromagnetic waves by using the reflections that occur at interfaces between dissimilar dielectric materials. They can be constructed for use at microwave frequencies, but are most commonly used at optical frequencies, where they can offer extremely low loss propagation. The most common dielectric waveguides are optical fibers, which are discussed elsewhere in this handbook (Chapter 14: Optical Communications).

Resonators are either metal or dielectric enclosures that exhibit sharp resonances at frequencies that can be controlled by choosing the size and material construction of the resonator. They are electromagnetic analogs of lumped resonant circuits and are typically used at microwave frequencies and above. Resonators can be constructed using a large variety of shaped enclosures, but simple shapes are usually chosen so that their resonant frequencies can be easily predicted and controlled. Typical shapes are rectangular and circular cylinders.

7.2. MODE CLASSIFICATIONS

Figure 7.1 shows a uniform waveguide, whose cross-sectional dimensions and material properties are constant along the waveguide (i.e., z) axis. Every type of waveguide has an

© 2006 by Taylor & Francis Group, LLC

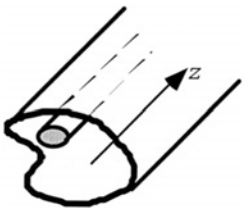


Figure 7.1 A uniform waveguide with arbitrary cross section.

infinite number of distinct electromagnetic field configurations that can exist inside it. Each of these configurations is called a *mode*. The characteristics of these modes depend upon the cross-sectional dimensions of the conducting cylinder, the type of dielectric material inside the waveguide, and the frequency of operation.

When waveguide properties are uniform along the z axis, the phasors representing the forward-propagating (i.e., $+z$) time-harmonic modes vary with the longitudinal coordinate z as $\mathbf{E}$, $\mathrm{H} \propto e^{-\gamma z}$, where the $e^{j\omega t}$ phasor convention is assumed. The parameter γ is called the *propagation constant* of the mode and is, in general, complex valued:

$$\gamma = \alpha + j\beta \tag{7.1}$$

where $j = \sqrt{-1}$, α is the modal attenuation constant, which controls the rate of decay of the wave amplitude, β is the phase constant, which controls the rate at which the phase of the wave changes, which in turn controls a number of other modal characteristics, including wavelength and velocity.

Waveguide modes are typically classed according to the nature of the electric and magnetic field components that are directed along the waveguide axis, E_z and H_z, which are called the *longitudinal components*. From Maxwell's equations, it follows that the *transverse components* (i.e., directed perpendicular to the direction of propagation) are related to the longitudinal components by the relations [1]

$$E_x = -\frac{1}{h^2}\left(\gamma\frac{\partial E_z}{\partial x} + j\omega\mu\frac{\partial H_z}{\partial y}\right) \tag{7.2}$$

$$E_y = -\frac{1}{h^2}\left(\gamma\frac{\partial E_z}{\partial y} - j\omega\mu\frac{\partial H_z}{\partial x}\right) \tag{7.3}$$

$$H_x = -\frac{1}{h^2}\left(-j\omega\varepsilon\frac{\partial E_z}{\partial y} + \gamma\frac{\partial H_z}{\partial x}\right) \tag{7.4}$$

$$H_y = -\frac{1}{h^2}\left(j\omega\varepsilon\frac{\partial E_z}{\partial x} + \gamma\frac{\partial H_z}{\partial y}\right) \tag{7.5}$$

where,

$$h^2 = k^2 + \gamma^2 \tag{7.6}$$

$k = 2\pi f\sqrt{\mu\varepsilon}$ is the *wave number* of the dielectric, $f = \omega/2\pi$ is the operating frequency in Hz, and μ and ε are the permeability and permittivity of the dielectric, respectively. Similar

© 2006 by Taylor & Francis Group, LLC

expressions for the transverse fields can be derived in other coordinate systems, but regardless of the coordinate system, the transverse fields are completely determined by the spatial derivatives of longitudinal field components across the cross section of the waveguide.

Several types of modes are possible in waveguides.

TE modes: Transverse-electric modes, sometimes called *H modes*. These modes have $E_z = 0$ at all points within the waveguide, which means that the electric field vector is always perpendicular (i.e., transverse) to the waveguide axis. These modes are always possible in metal waveguides with homogeneous dielectrics.

TM modes: Transverse-magnetic modes, sometimes called *E modes*. These modes have $H_z = 0$ at all points within the waveguide, which means that the magnetic field vector is perpendicular to the waveguide axis. Like TE modes, they are always possible in metal waveguides with homogeneous dielectrics.

EH modes: These are hybrid modes in which neither E_z nor H_z is zero, but the characteristics of the transverse fields are controlled more by E_z than H_z. These modes usually occur in dielectric waveguides and metal waveguides with inhomogeneous dielectrics.

HE modes: These are hybrid modes in which neither E_z nor H_z is zero, but the characteristics of the transverse fields are controlled more by H_z than E_z. Like EH modes, these modes usually occur in dielectric waveguides and in metal waveguides with inhomogeneous dielectrics.

TEM modes: Transverse-electromagnetic modes, often called *transmission-line modes*. These modes can exist only when more than one conductor with a complete dc circuit path is present in the waveguide, such as the inner and outer conductors of a coaxial cable. These modes are not considered to be waveguide modes.

Both transmission lines and waveguides are capable of guiding electromagnetic signal energy over long distances, but waveguide modes behave quite differently with changes in frequency than do transmission-line modes. The most important difference is that waveguide modes can typically transport energy only at frequencies above distinct cutoff frequencies, whereas transmission line modes can transport energy at frequencies all the way down to dc. For this reason, the term *transmission line* is reserved for structures capable of supporting TEM modes, whereas the term *waveguide* is typically reserved for structures that can only support waveguide modes.

7.3. MODAL FIELDS AND CUTOFF FREQUENCIES

For all uniform waveguides, E_z and H_z satisfy the scalar wave equation at all points within the waveguide [1]:

$$\nabla^2 E_z + k^2 E_z = 0 \tag{7.7}$$

$$\nabla^2 H_z + k^2 H_z = 0 \tag{7.8}$$

where ∇^2 is the Laplacian operator and k is the wave number of the dielectric. However, for $+z$ propagating fields, $\partial()/\partial z = -\gamma()$, so we can write

$$\nabla_t^2 E_z + h^2 E_z = 0 \tag{7.9}$$

© 2006 by Taylor & Francis Group, LLC

and

$$\nabla_t^2 H_z + h^2 H_z = 0 \tag{7.10}$$

where h^2 is given by Eq. (7.5) and ∇_t^2 is the transverse Laplacian operator. In Cartesian coordinates, $\nabla_t^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$. When more than one dielectric is present, E_z and H_z must satisfy Eqs. (7.9) and (7.10) in each region for the appropriate value of k in each region.

Modal solutions are obtained by first finding general solutions to Eqs. (7.9) and (7.10) and then applying boundary conditions that are appropriate for the particular waveguide. In the case of metal waveguides, $E_z = 0$ and $\partial H_z/\partial p = 0$ at the metal walls, where p is the direction perpendicular to the waveguide wall. At dielectric–dielectric interfaces, the E- and H-field components tangent to the interfaces must be continuous. Solutions exist for only certain values of h, called *modal eigenvalues*. For metal waveguides with homogeneous dielectrics, each mode has a single modal eigenvalue, whose value is independent of frequency. Waveguides with multiple dielectrics, on the other hand, have different modal eigenvalues in each dielectric region and are functions of frequency, but the propagation constant γ is the same in each region.

Regardless of the type of waveguide, the propagation constant γ for each mode is determined by its modal eigenvalue, the frequency of operation, and the dielectric properties. From Eqs. (7.1) and (7.6), it follows that

$$\gamma = \alpha + j\beta = \sqrt{h^2 - k^2} \tag{7.11}$$

where h is the modal eigenvalue associated with the dielectric wave number k. When a waveguide has no material or radiation (i.e., leakage) loss, the modal eigenvalues are always real-valued. For this case, γ is either real or imaginary. When $k^2 > h^2$, $\alpha = 0$ and $\beta > 0$, so the modal fields are propagating fields with no attenuation. On the other hand, when $k^2 < h^2$, $\alpha > 0$, and $\beta = 0$, which means that the modal fields are nonpropagating and decay exponentially with distance. Fields of this type are called *evanescent fields*. The frequency at which $k^2 = h^2$ is called the *modal cutoff frequency* f_c. A mode operated at frequencies above its cutoff frequency is a propagating mode. Conversely, a mode operated below its cutoff frequency is an evanescent mode.

The dominant mode of a waveguide is the one with the lowest cutoff frequency. Although higher order modes are often useful for a variety of specialized uses of waveguides, signal distortion is usually minimized when a waveguide is operated in the frequency range where only the dominant mode is propagating. This range of frequencies is called the *dominant rang*e of the waveguide.

7.4. PROPERTIES OF METAL WAVEGUIDES

Metal waveguides are the most commonly used waveguides at RF and microwave frequencies. Like coaxial transmission lines, they confine fields within a conducting shell, which reduces cross talk with other circuits. In addition, metal waveguides usually exhibit lower losses than coaxial transmission lines of the same size. Although they can be constructed using more than one dielectric, most metal waveguides are simply metal pipes filled with a homogeneous dielectric—usually air. In the remainder of this chapter, the term *metal waveguides* will denote self-enclosed metal waveguides with homogeneous dielectrics.

© 2006 by Taylor & Francis Group, LLC

Metal waveguides have the simplest electrical characteristics of all waveguide types, since their modal eigenvalues are functions only of the cross-sectional shape of the metal cylinder and are independent of frequency. For this case, the amplitude and phase constants of any allowed mode can be written in the form:

$$\alpha = \begin{cases} h\sqrt{1-\left(\dfrac{f}{f_c}\right)^2} & \text{for } f < f_c \\ 0 & \text{for } f > f_c \end{cases} \tag{7.12}$$

and

$$\beta = \begin{cases} 0 & \text{for } f < f_c \\ h\sqrt{\left(\dfrac{f}{f_c}\right)^2 - 1} & \text{for } f > f_c \end{cases} \tag{7.13}$$

where

$$f_c = \frac{h}{2\pi\sqrt{\mu\varepsilon}} \tag{7.14}$$

Each mode has a unique modal eigenvalue h, so each mode has a specific cutoff frequency. The mode with the smallest modal eigenvalue is the dominant mode. If two or more modes have the same eigenvalue, they are *degenerate modes*.

7.4.1. Guide Wavelength

The distance over which the phase of a propagating mode in a waveguide advances by 2π is called the *guide wavelength* λ_g. For metal waveguides, β is given by Eq. (7.13), so λ_g for any mode can be expressed as

$$\lambda_g = \frac{2\pi}{\beta} = \frac{\lambda}{\sqrt{1-(f_c/f)^2}} \tag{7.15}$$

where $\lambda = (f\sqrt{\mu\varepsilon})^{-1}$ is the wavelength of a plane wave of the same frequency in the waveguide dielectric. For $f \gg f_c$, $\lambda_g \approx \lambda$. Also, $\lambda_g \to \infty$ as $f \to f_c$, which is one reason why it is usually undesirable to operate a waveguide mode near modal cutoff frequencies.

7.4.2. Wave Impedance

Although waveguide modes are not plane waves, the ratio of their transverse electric and magnetic field magnitudes are constant throughout the cross sections of the metal waveguides, just as for plane waves. This ratio is called the modal *wave impedance* and has the following values for TE and TM modes [1]:

$$Z_{\text{TE}} = \frac{E_T}{H_T} = \frac{j\omega\mu}{\gamma} = \frac{\eta}{\sqrt{1-(f_c/f)^2}} \tag{7.16}$$

© 2006 by Taylor & Francis Group, LLC

and

$$Z_{\mathrm{TM}} = \frac{E_T}{H_T} = \frac{\gamma}{j\omega\varepsilon} = \eta\sqrt{1-(f_c/f)^2} \tag{7.17}$$

where E_T and H_T are the magnitudes of the transverse electric and magnetic fields, respectively, and $\eta = \sqrt{\mu/\varepsilon}$ is the intrinsic impedance of the dielectric. In the limit as $f \to \infty$, both Z_{TE} and Z_{TM} approach η. On the other hand, as $f \to f_c$, $Z_{\mathrm{TE}} \to \infty$ and $Z_{\mathrm{TM}} \to 0$, which means that the transverse electric fields are dominant in TE modes near cutoff and the transverse magnetic fields are dominant in TM modes near cutoff.

7.4.3. Wave Velocities

The phase and group velocities of waveguide modes are both related to the rates of change of the modal propagation constant β with respect to frequency. The phase velocity u_p is the velocity of the phase fronts of the mode along the waveguide axis and is given by [1]

$$u_p = \frac{\omega}{\beta} \tag{7.18}$$

Conversely, the group velocity is the velocity at which the amplitude envelopes of narrow-band, modulated signals propagate and is given by [1]

$$u_g = \frac{\partial\omega}{\partial\beta} = \left(\frac{\partial\beta}{\partial\omega}\right)^{-1} \tag{7.19}$$

Unlike transmission-line modes, where β is a linear function frequency, β is not a linear function of frequency for waveguide modes; so u_p and u_g are not the same for waveguide modes. For metal waveguides, it is found from Eqs. (7.13), (7.18), and (7.19) that

$$u_p = \frac{u_{\mathrm{TEM}}}{\sqrt{1-(f_c/f)^2}} \tag{7.20}$$

and

$$u_g = u_{\mathrm{TEM}}\sqrt{1-(f_c/f)^2} \tag{7.21}$$

where $u_{\mathrm{TEM}} = 1/\sqrt{\mu\varepsilon}$ is the velocity of a plane wave in the dielectric.

Both u_p and u_g approach u_{TEM} as $f \to \infty$, which is an indication that waveguide modes appear more and more like TEM modes at high frequencies. But near cutoff, their behaviors are very different: u_g approaches zero, whereas u_p approaches infinity. This behavior of u_p may at first seem at odds with Einstein's theory of special relativity, which states that energy and matter cannot travel faster than the vacuum speed of light c. But this result is not a violation of Einstein's theory since neither information nor energy is conveyed by the phase of a steady-state waveform. Rather, the energy and information are transported at the group velocity, which is always less than or equal to c.

© 2006 by Taylor & Francis Group, LLC

7.4.4. Dispersion

Unlike the modes on transmission lines, which exhibit differential propagation delays (i.e., *dispersion*) only when the materials are lossy or frequency dependent, waveguide modes are always dispersive, even when the dielectric is lossless and walls are perfectly conducting. The pulse spread per meter Δt experienced by a modulated pulse is equal to the difference between the arrival times of the lowest and highest frequency portions of the pulse. Since the envelope delay per meter for each narrow-band components of a pulse is equal to the inverse of the group velocity at that frequency, we find that the pulse spreading Δt for the entire pulse is given by

$$\Delta t = \left.\frac{1}{u_g}\right|_{\max} - \left.\frac{1}{u_g}\right|_{\min} \tag{7.22}$$

where $1/u_g|_{\max}$ and $1/u_g|_{\min}$ are the maximum and minimum inverse group velocities encountered within the pulse bandwidth, respectively. Using Eq. (7.21), the pulse spreading in metal waveguides can be written as

$$\Delta t = \frac{1}{u_{\text{TEM}}}\left(\frac{1}{\sqrt{1-(f_c/f_{\min})^2}} - \frac{1}{\sqrt{1-(f_c/f_{\max})^2}}\right) \tag{7.23}$$

where $f_{\min}$ and $f_{\max}$ are the minimum and maximum frequencies within the pulse 3-dB bandwidth. From this expression, it is apparent that pulse broadening is most pronounced when a waveguide mode is operated close to its cutoff frequency f_c.

The pulse spreading specified by Eq. (7.23) is the result of *waveguide dispersion*, which is produced solely by the confinement of a wave by a guiding structure and has nothing to do with any frequency-dependent parameters of the waveguide materials. Other dispersive effects in waveguides are *material dispersion* and *modal dispersion*. Material dispersion is the result of frequency-dependent characteristics of the materials used in the waveguide, usually the dielectric. Typically, material dispersion causes higher frequencies to propagate more slowly than lower frequencies. This is often termed *normal dispersion*. Waveguide dispersion, on the other hand, causes the opposite effect and is often termed *anomalous dispersion*.

Modal dispersion is the spreading that occurs when the signal energy is carried by more than one waveguide mode. Since each mode has a distinct group velocity, the effects of modal dispersion can be very severe. However, unlike waveguide dispersion, modal dispersion can be eliminated simply by insuring that a waveguide is operated only in its dominant frequency range.

7.4.5. Effects of Losses

There are two mechanisms that cause losses in metal waveguides: dielectric losses and metal losses. In both cases, these losses cause the amplitudes of the propagating modes to decay as $e^{-\alpha z}$, where α is the attenuation constant, measured in units of Nepers per meter. Typically, the attenuation constant is considered as the sum of two components: $\alpha = \alpha_d + \alpha_c$, where α_d and α_c are the attenuation constants due to dielectric and metal losses alone, respectively. In most cases, dielectric losses are negligible compared to metal losses, in which case $\alpha \approx \alpha_c$.

© 2006 by Taylor & Francis Group, LLC

Often, it is useful to specify the attenuation constant of a mode in terms of its decibel loss per meter length, rather than in Nepers per meter. The conversion formula between the two unit conventions is

$$\alpha \text{ (dB/m)} = 8.686 \times \alpha \text{ (Np/m)} \tag{7.24}$$

Both unit systems are useful, but it should be noted that α must be specified in Np/m when it is used in formulas that contain the terms of the form $e^{-\alpha z}$.

The attenuation constant α_d can be found directly from Eq. (7.11) simply by generalizing the dielectric wave number k to include the effect of the dielectric conductivity σ. For a lossy dielectric, the wave number is given by $k^2 = \omega^2\mu\varepsilon(1 + \sigma/j\omega\varepsilon)$, where σ is the conductivity of the dielectric, so the attenuation constant α_d due to dielectric losses alone is given by

$$\alpha_d = \text{Re}\left(\sqrt{h^2 - \omega^2\mu\varepsilon\left(1+\frac{\sigma}{j\omega\varepsilon}\right)}\right) \tag{7.25}$$

where Re signifies "the real part of" and h is the modal eigenvalue.

The effect of metal loss is that the tangential electric fields at the conductor boundary are no longer zero. This means that the modal fields exist both in the dielectric and the metal walls. Exact solutions for this case are much more complicated than the lossless case. Fortunately, a perturbational approach can be used when wall conductivities are high, as is usually the case. For this case, the modal field distributions over the cross section of the waveguide are disturbed only slightly; so a perturbational approach can be used to estimate the metal losses except at frequencies very close to the modal cutoff frequency [2].

This perturbational approach starts by noting that the power transmitted by a waveguide mode decays as

$$P = P_0 e^{-2\alpha_c z} \tag{7.26}$$

where P_0 is the power at $z = 0$. Differentiating this expression with respect to z, solving for α_c, and noting that dP/dz is the negative of the power loss per meter P_L, it is found that

$$a_c = \frac{1}{2}\frac{P_L}{P} \tag{7.27}$$

Expressions for α_c in terms of the modal fields can be found by first recognizing that the transmitted power P is integral of the average Poynting vector over the cross section S of the waveguide [1]:

$$P = \frac{1}{2}\text{Re}\left(\int_S \mathbf{E} \times \mathbf{H}^* \cdot \mathbf{ds}\right) \tag{7.28}$$

where "*" indicates the complex conjugate, and "$\cdot$" and "$\times$" indicate the dot and cross products, respectively.

© 2006 by Taylor & Francis Group, LLC

Similarly, the power loss per meter can be estimated by noting that the wall currents are controlled by the tangential H field at the conducting walls. When conductivities are high, the wall currents can be treated as if they flow uniformly within a skin depth of the surface. The resulting expression can be expressed as [1]

$$P_L = \frac{1}{2} R_s \oint_C |H|^2 dl \tag{7.29}$$

where $R_s = \sqrt{\pi f \mu/\sigma}$ is the surface resistance of the walls (μ and σ are the permeability and conductivity of the metal walls, respectively) and the integration takes place along the perimeter of the waveguide cross section.

As long as the metal losses are small and the operation frequency is not too close to cuttoff, the modal fields for the perfectly conducting case can be used in the above integral expressions for P and P_L. Closed form expressions for α_c for rectangular and circular waveguide modes are presented later in this chapter.

7.5. RECTANGULAR WAVEGUIDES

A rectangular waveguide is shown in Fig. 7.2, consisting of a rectangular metal cylinder of width a and height b, filled with a homogenous dielectric with permeability and permittivity μ and ε, respectively. By convention, it is assumed that $a \geq b$. If the walls are perfectly conducting, the field components for the TE_{mn} modes are given by

$$E_x = H_0 \frac{j\omega\mu}{h_{mn}^2} \frac{n\pi}{b} \cos\left(\frac{m\pi}{a}x\right) \sin\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30a}$$

$$E_y = -H_0 \frac{j\omega\mu}{h_{mn}^2} \frac{m\pi}{a} \sin\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30b}$$

$$E_z = 0 \tag{7.30c}$$

$$H_x = H_0 \frac{\gamma_{mn}}{h_{mn}^2} \frac{m\pi}{a} \sin\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30d}$$

$$H_y = H_0 \frac{\gamma_{mn}}{h_{mn}^2} \frac{n\pi}{b} \cos\left(\frac{m\pi}{a}x\right) \sin\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30e}$$

$$H_z = H_0 \cos\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30f}$$

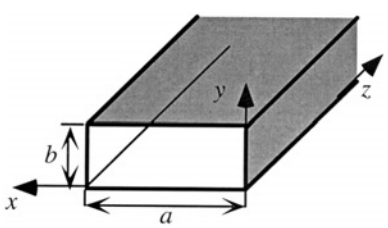


Figure 7.2 A rectangular waveguide.

© 2006 by Taylor & Francis Group, LLC

The modal eigenvalues, propagation constants, and cutoff frequencies are

$$h_{mn} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} \tag{7.31}$$

$$\gamma_{mn} = \alpha_{mn} + j\beta_{mn} = j(2\pi f)\sqrt{\mu\varepsilon}\sqrt{1 - \left(\frac{f_{c_{mn}}}{f}\right)^2} \tag{7.32}$$

$$f_{c_{mn}} = \frac{1}{2\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2} \tag{7.33}$$

For the TE_{mn} modes, m and n can be any positive integer values, including zero, so long as both are not zero.

The field components for the TM_{mn} modes are

$$E_x = -E_0\left(\frac{\gamma_{mn}}{h_{mn}^2}\right)\left(\frac{m\pi}{a}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34a}$$

$$E_y = -E_0\frac{\gamma_{mn}}{h_{mn}^2}\frac{n\pi}{b}\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34b}$$

$$E_z = E_0\sin\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34c}$$

$$H_x = E_0\frac{j\omega\varepsilon}{h_{mn}^2}\frac{n\pi}{b}\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34d}$$

$$H_y = -E_0\frac{j\omega\varepsilon}{h_{mn}^2}\frac{m\pi}{a}\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34e}$$

$$H_z = 0 \tag{7.34f}$$

where the values of h_{mn}, γ_{mn}, and $f_{c_{mn}}$ are the same as for the TE_{mn} modes [Eqs. (7.31)–(7.33)]. For the TM_{mn} modes, m and n can be any positive integer value except zero.

The dominant mode in a rectangular waveguide is the TE_{10} mode, which has a cutoff frequency of

$$f_{c_{10}} = \frac{1}{2a\sqrt{\mu\varepsilon}} \tag{7.35}$$

The modal field patterns for this mode are shown in Fig. 7.3. Table 7.1 shows the cutoff frequencies of the lowest order rectangular waveguide modes (referenced to the

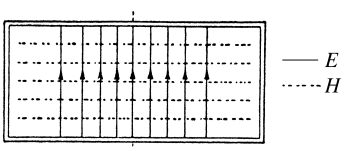


Figure 7.3 Field configuration for the TE_{10} (dominant) mode of a rectangular waveguide. (Adapted from Ref. 2 with permission.)

© 2006 by Taylor & Francis Group, LLC

Table 7.1 Cutoff Frequencies of the Lowest Order Rectangular Waveguide Modes for $a/b = 2.1$.

f_c/f_{c10}	Modes
1.0	TE_{10}
2.0	TE_{20}
2.1	TE_{01}
2.326	TE_{11}, TM_{11}
2.9	TE_{21}, TM_{21}
3.0	TE_{30}
3.662	TE_{31}, TM_{31}
4.0	TE_{40}

Frequencies are Referenced to the Cutoff Frequency of the Dominant Mode.

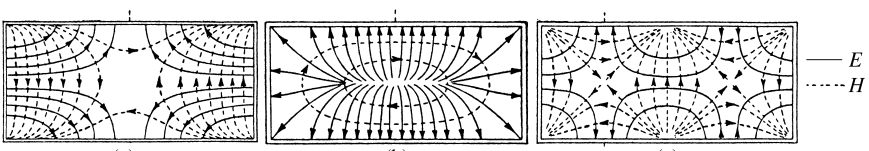


Figure 7.4 Field configurations for the TE_{11}, TM_{11}, and the TE_{21} modes in rectangular waveguides. (Adapted from Ref. 2 with permission.)

cutoff frequency of the dominant mode) when $a/b = 2.1$. The modal field patterns of several lower order modes are shown in Fig. 7.4.

The attenuation constants that result from metal losses alone can be obtained by substituting the modal fields into Eqs. (7.27)–(7.29). The resulting expressions are [3]

$$\alpha_{mn} = \frac{2R_s}{b\eta\left(1 - h_{mn}^2/k^2\right)^{1/2}}\left[\frac{h_{mn}^2}{k^2}\left(1 + \frac{b}{a}\right) + \frac{b}{a}\left(\frac{\varepsilon_{0m}}{2} - \frac{h_{mn}^2}{k^2}\right)\left(\frac{n^2ab + m^2a^2}{n^2b^2 + m^2a^2}\right)\right] \qquad \text{TE modes} \tag{7.36}$$

and

$$\alpha_{mn} = \frac{2R_s}{b\eta\left(1 - h_{mn}^2/k^2\right)^{1/2}}\left(\frac{n^2b^3 + m^2a^3}{n^2b^2a + m^2a^3}\right) \qquad \text{TM modes} \tag{7.37}$$

where $R_s = \sqrt{\pi f \mu/\sigma}$ is the surface resistance of the metal, η is the intrinsic impedance of the dielectric (377 Ω for air), $\varepsilon_{0m} = 1$ for $m = 0$ and 2 for $m > 0$, and the modal eigenvalues h_{mn} are given by Eq. (7.31). Figure 7.5 shows the attenuation constant for several lower order modes as a function of frequency. In each case, losses are highest at frequencies near the modal cutoff frequencies.

© 2006 by Taylor & Francis Group, LLC

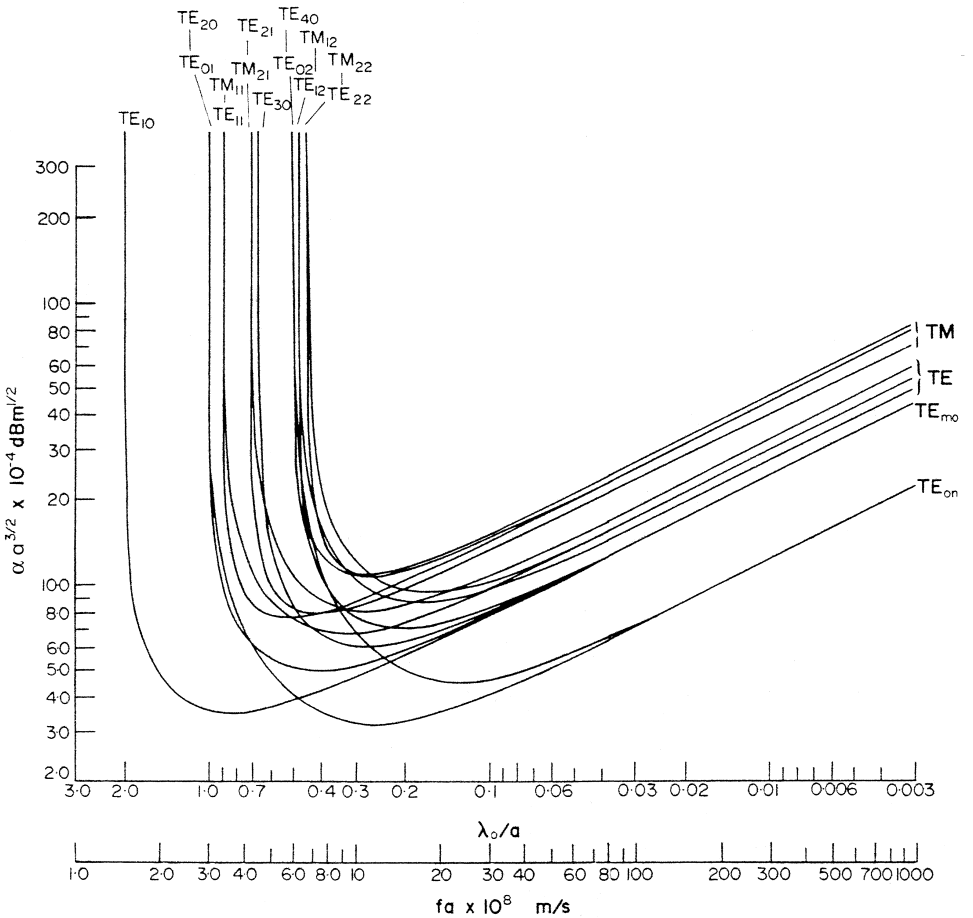


Figure 7.5 The attenuation constant of several lower order modes due to metal losses in rectangular waveguides with $a/b = 2$, plotted against normalized wavelength. (Adapted from Baden Fuller, A.J. *Microwaves*, 2nd Ed.; Oxford: Pergamon Press Ltd., 1979, with permission.)

7.6. CIRCULAR WAVEGUIDES

A circular waveguide with inner radius a is shown in Fig. 7.6, consisting of a rectangular metal cylinder with inside radius a, filled with a homogenous dielectric. The axis of the waveguide is aligned with the z axis of a circular-cylindrical coordinate system, where ρ and ϕ are the radial and azimuthal coordinates, respectively. If the walls are perfectly conducting, the equations for the TE_{nm} modes are

$$E_\rho = H_0 \frac{j\omega\mu n}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38a}$$

$$E_\phi = H_0 \frac{j\omega\mu}{h_{nm}} J_n'(h_{nm}\rho) \cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38b}$$

$$E_z = 0 \tag{7.38c}$$

© 2006 by Taylor & Francis Group, LLC

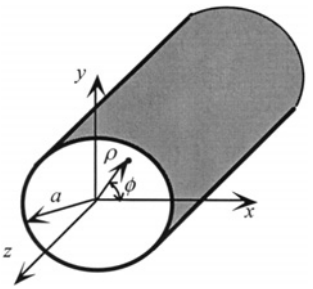


Figure 7.6 A circular waveguide.

$$H_\rho = -H_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38d}$$

$$H_\phi = H_0 \frac{\gamma_{nm} n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38e}$$

$$H_z = H_0 J_n(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38f}$$

where n is any positive valued integer, including zero and $J_n(x)$ and $J_n'(x)$ are the regular Bessel function of order n and its first derivative [4,5], respectively, and μ and ε are the permeability and permittivity of the interior dielectric, respectively. The allowed modal eigenvalues h_{nm} are

$$h_{mn} = \frac{p'_{nm}}{a} \tag{7.39}$$

Here, the values p'_{nm} are roots of the equation

$$J_n'(p'_{nm}) = 0 \tag{7.40}$$

where m signifies the mth root of $J_n'(x)$. By convention, $1 < m < \infty$, where $m = 1$ indicates the smallest root. Also for the TE modes,

$$\gamma_{nm} = \alpha_{nm} + j\beta_{nm} = j(2\pi f)\sqrt{\mu\varepsilon}\sqrt{1 - \left(\frac{f_{c_{nm}}}{f}\right)^2} \tag{7.41}$$

$$f_{c_{nm}} = \frac{p'_{nm}}{2\pi a\sqrt{\mu\varepsilon}} \tag{7.42}$$

The equations that define the TM_{nm} modes in circular waveguides are

$$E_\rho = -E_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.43a}$$

$$E_\phi = E_0 \frac{\gamma_{nm} n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.43b}$$

© 2006 by Taylor & Francis Group, LLC

$$E_z = E_0 J_n(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.43c}$$

$$H_\rho = -E_0 \frac{j\omega\varepsilon n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.43d}$$

$$H_\phi = -E_0 \frac{j\omega\varepsilon}{h_{nm}} J_n'(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.43e}$$

$$H_z = 0 \tag{7.43f}$$

where n is any positive valued integer, including zero. For the TM_{nm} modes, the values of the modal eigenvalues are given by

$$h_{nm} = \frac{p_{nm}}{a} \tag{7.44}$$

Here, the values p_{nm} are roots of the equation

$$J_n(p_{nm}) = 0 \tag{7.45}$$

where m signifies the mth root of $J_n(x)$, where $1 < m < \infty$. Also for the TM modes,

$$\gamma_{nm} = \alpha_{nm} + j\beta_{nm} = j(2\pi f)\sqrt{\mu\varepsilon}\sqrt{1 - \left(\frac{f_{c_{mn}}}{f}\right)^2} \tag{7.46}$$

$$f_{c_{nm}} = \frac{p_{nm}}{2\pi a\sqrt{\mu\varepsilon}} \tag{7.47}$$

The dominant mode in a circular waveguide is the TE_{11} mode, which has a cutoff frequency given by

$$f_{c_{11}} = \frac{0.293}{a\sqrt{\mu\varepsilon}} \tag{7.48}$$

The configurations of the electric and magnetic fields of this mode are shown in Fig. 7.7. Table 7.2 shows the cutoff frequencies of the lowest order modes for circular waveguides, referenced to the cutoff frequency of the dominant mode. The modal field patterns of several lower order modes are shown in Fig. 7.8.

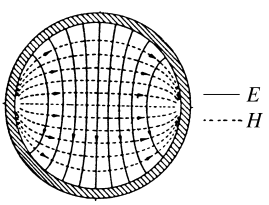


Figure 7.7 Field configuration for the TE_{11} (dominant) mode in a circular waveguide. (Adapted from Ref. 2 with permission.)

© 2006 by Taylor & Francis Group, LLC

Table 7.2 Cutoff Frequencies of the Lowest Order Circular Waveguide Modes.

f_c/f_{c11}	Modes
1.0	TE_{11}
1.307	TM_{01}
1.66	TE_{21}
2.083	TE_{01}, TM_{11}
2.283	TE_{31}
2.791	TM_{21}
2.89	TE_{41}
3.0	TE_{12}

Frequencies are Referenced to the Cutoff Frequency of the Dominant Mode.

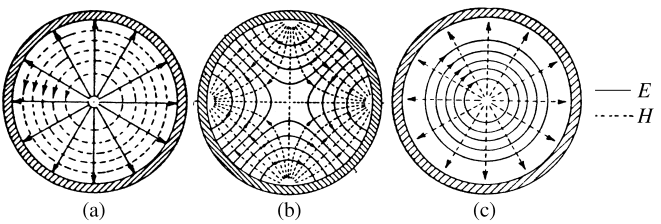


Figure 7.8 Field configurations of the TM_{01}, TE_{01}, and TE_{21} modes in a circular waveguide. (Adapted from Ref. 2 with permission.)

The attenuation constants that result from metal losses alone can be obtained by substituting the modal fields into Eqs. (7.27)–(7.29). The resulting expressions are [3]

$$\alpha_{nm} = \frac{R_s}{a\eta\left[1-(p'_{nm}/ka)^2\right]^{1/2}}\left[\frac{(p'_{nm})^2}{a^2k^2}+\frac{n^2}{(p'_{nm})^2-n^2}\right] \qquad \text{TE modes} \tag{7.49}$$

and

$$\alpha_{nm} = \frac{R_s}{a\eta\left[1-(p_{nm}/ka)^2\right]^{1/2}} \qquad \text{TM modes} \tag{7.50}$$

Figure 7.9 shows the metal attenuation constants for several circular waveguide modes, each normalized to the surface resistance R_s of the walls. As can be seen from this figure, the TE_{0m} modes exhibit particularly low loss at frequencies significantly above their cutoff frequencies, making them useful for transporting microwave energy over large distances.

7.7. COAXIAL-TO-WAVEGUIDE TRANSITIONS

When coupling electromagnetic energy into a waveguide, it is important to ensure that the desired mode is excited and that reflections back to the source are minimized, and

© 2006 by Taylor & Francis Group, LLC

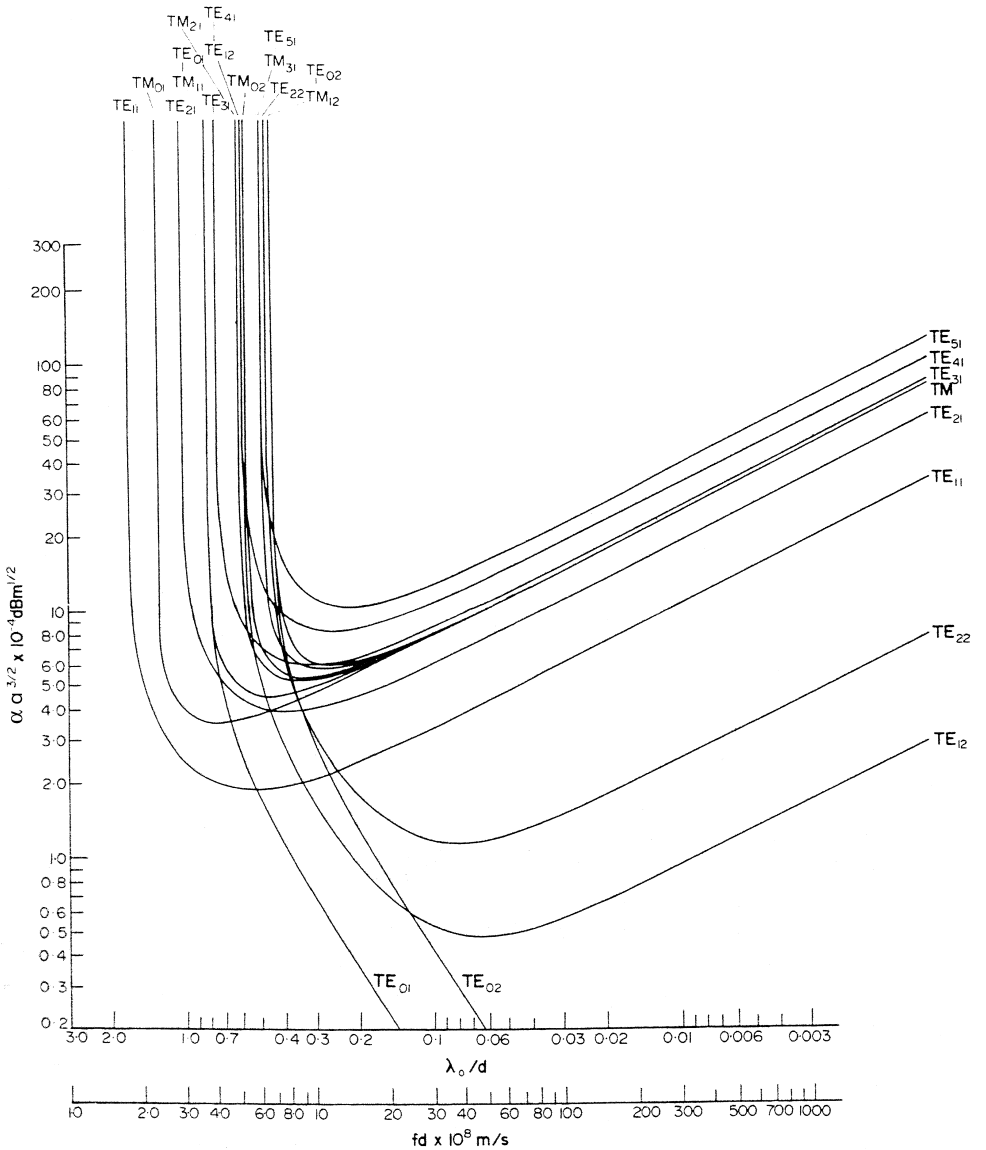


Figure 7.9 The attenuation constant of several lower order modes due to metal losses in circular waveguides with diameter d, plotted against normalized wavelength. (Adapted from Baden Fuller, A.J. *Microwaves*, 2nd Ed.; Oxford: Pergamon Press Ltd., 1979, with permission.)

that undesired higher order modes are not excited. Similar concerns must be considered when coupling energy from a waveguide to a transmission line or circuit element. This is achieved by using launching (or coupling) structures that allow strong coupling between the desired modes on both structures.

Figure 7.10 shows a mode launching structure launching the TE_{10} mode in a rectangular waveguide from a coaxial transmission line. This structure provides good coupling between the TEM (transmission line) mode on the coaxial line and the

© 2006 by Taylor & Francis Group, LLC

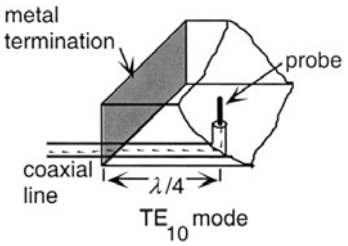


Figure 7.10 Coaxial-to-rectangular waveguide transition that couples the coaxial line to the TE_{10} waveguide mode.

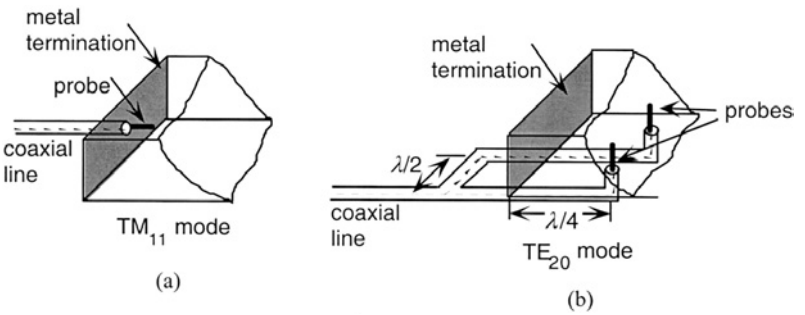


Figure 7.11 Coaxial-to-rectangular transitions that excite the TM_{11} and TM_{12} modes.

TE_{10} mode. The probe extending from inner conductor of the coaxial line excites a strong vertical electric field in the center of the waveguide, which matches the TE_{10} modal E field. The distance between the probe and the short circuit back wall is chosen to be approximately $\lambda_g/4$, which allows the backward-launched fields to reflect off the short circuit and arrive in phase with the fields launched toward the right.

Launching structures can also be devised to launch higher order modes. Mode launchers that couple the transmission line mode on a coaxial cable to the TM_{11} and TE_{20} waveguide modes are shown in Fig. 7.11.

7.8. COMPARATIVE SURVEY OF METAL WAVEGUIDES

All waveguides are alike in that they can propagate electromagnetic signal energy via an infinite number of distinct waveguide modes. Even so, each waveguide type has certain specific electrical or mechanical characteristics that may make it more or less suitable for a specific application. This section briefly compares the most notable features of the most common types: rectangular, circular, elliptical, and ridge waveguides.

Rectangular waveguides are popular because they have a relatively large dominant range and moderate losses. Also, since the cutoff frequencies of the TE_{10} and TE_{01} modes are different, it is impossible for the polarization direction to change when a rectangular waveguide is operated in its dominant range, even when nonuniformities such as bends and obstacles are encountered. This is important when feeding devices such as antennas, where the polarization of the incident field is critical.

© 2006 by Taylor & Francis Group, LLC

Circular waveguides have a smaller dominant range than rectangular waveguides. While this can be a disadvantage, circular waveguides have several attractive features. One of them is their shape, which allows the use of circular terminations and connectors, which are easier to manufacture and attach. Also, circular waveguides maintain their shapes reasonably well when they are bent, so they can be easily routed between the components of a system. Circular waveguides are also used for making rotary joints, which are needed when a section of waveguide must be able to rotate, such as for the feeds of revolving antennas. Another useful characteristic of circular waveguides is that some of their higher order modes have particularly low loss. This makes them attractive when signals must be sent over relatively long distances, such as for the feeds of microwave antennas on tall towers.

An elliptical waveguide is shown in Fig. 7.12a. As might be expected by their shape, elliptical waveguides bear similarities to both circular and rectangular waveguides. Like circular waveguides, they are easy to bend. The modes of elliptical waveguides can be expressed in terms of Mathieu functions [6] and are similar to those of circular waveguides, but exhibit different cutoff frequencies for modes polarized along the major and minor axes of the elliptical cross section of the waveguide. This means that unlike circular waveguides, where the direction of polarization tends to rotate as the waves pass through bends and twists, modal polarization is much more stable in elliptical waveguides. This property makes elliptical waveguides attractive for feeding certain types of antennas, where the polarization state at the input to the antenna is critical.

Single and double ridge waveguides are shown in Fig. 7.12b and c, respectively. The modes of these waveguides bear similarities to those of rectangular guides, but can only be derived numerically [7]. Nevertheless, the effect of the ridges can be seen by realizing that they act as a uniform, distributed capacitance that reduces the characteristic impedance of the waveguide and lowers its phase velocity. This reduced phase velocity results in a lowering of the cutoff frequency of the dominant mode by a factor of 5 or higher, depending upon the dimensions of the ridges. Thus, the dominant range of a ridge waveguide is much greater than that of a standard rectangular waveguide. However, this increased frequency bandwidth is obtained at the expense of increased loss and decreased power handling capacity. The increased loss occurs because of the concentration of current flow on the ridges, with result in correspondingly high ohmic losses. The decreased power handling capability is a result of increased E-field levels in the vicinity of the ridges, which can cause breakdown (i.e., arcing) in the dielectric.

Waveguides are also available in a number of constructions, including rigid, semi-rigid, and flexible. In applications where it is not necessary for the waveguide to bend, rigid construction is always the best since it exhibits the lowest loss. In general, the more flexible the waveguide construction, the higher the loss.

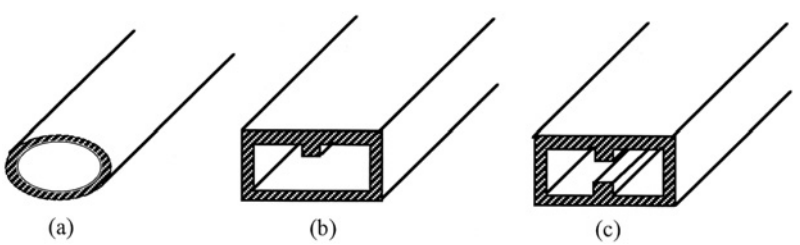


Figure 7.12 (a) Elliptical, (b) single-ridge, and (c) double-ridge waveguides.

© 2006 by Taylor & Francis Group, LLC

7.9. CAVITY RESONATORS

Resonant circuits are used for a variety of applications, including oscillator circuits, filters and tuned amplifiers. These circuits are usually constructed using lumped reactive components at audio through RF frequencies, but lumped components become less desirable at microwave frequencies and above. This is because at these frequencies, lumped components either do not exist or they are too lossy.

A more attractive approach at microwave frequencies and above is to construct devices that use the constructive and destructive interferences of multiply reflected waves to cause resonances. These reflections occur in enclosures called *cavity resonators*. Metal cavity resonators consist of metallic enclosures, filled with a dielectric (possibly air). Dielectric resonators are simply a solid block of dielectric material, surrounded by air. Cavity resonators are similar to waveguides in that they both support a large number of distinct modes. However, resonator modes are usually restricted to very narrow frequency ranges, whereas each waveguide mode can exist over a broad range of frequencies.

7.9.1. Cylindrical Cavity Resonators

A cylindrical cavity resonator is shown in Fig. 7.13, consisting of a hollow metal cylinder of radius a and length d, with metal end caps. The resonator fields can be considered to be combinations of upward- and downward-propagating waveguide modes. If the dielectric inside the resonator is homogeneous and the conducting walls are lossless, the TE fields are

$$E_\rho = H_0 \frac{j\omega\mu n}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin n\phi \left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51a}$$

$$E_\phi = H_0 \frac{j\omega\mu}{h_{nm}} J_n'(h_{nm}\rho) \cos n\phi \left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51b}$$

$$H_\rho = -H_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho) \cos n\phi \left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51c}$$

$$H_\phi = H_0 \frac{\gamma_{nm} n}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin n\phi \left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51d}$$

$$H_z = H_0 J_n(h_{nm}\rho) \cos n\phi \left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51e}$$

Here, the modal eigenvalues are $h_{nm} = p'_{nm}/a$, where the values of p'_{nm} are given by Eq. (7.40). To insure that E_ρ and E_ϕ vanish at $z = \pm d/2$, it is required that $A^- = A^+$ (even

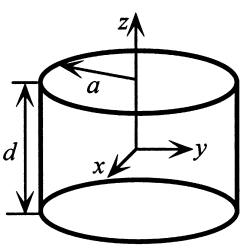


Figure 7.13 A cylindrical cavity resonator.

© 2006 by Taylor & Francis Group, LLC

modes) or $A^- = -A^+$ (odd modes) and that β_{nm} be restricted to the values $l\pi/d$, where $l = 0, 1, \ldots$. Each value of l corresponds to a unique frequency, called a *resonant frequency*. The resonant frequencies of the TE_{nml} modes are

$$f_{nml} = \frac{1}{2\pi\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{p'_{nm}}{a}\right)^2 + \left(\frac{l\pi}{d}\right)^2} \qquad (TE_{nml} \text{ modes}) \tag{7.52}$$

In a similar manner, the TM fields inside the resonator are of the form

$$E_\rho = -E_0 \frac{\gamma_{nm}}{h_{nm}} J'_n(h_{nm}\rho)\cos n\phi \left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.53a}$$

$$E_\phi = E_0 \frac{\gamma_{nm} n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi \left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.53b}$$

$$E_z = E_0 J_n(h_{nm}\rho)\cos n\phi \left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.53c}$$

$$H_\rho = -E_0 \frac{j\omega\varepsilon n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi \left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.53d}$$

$$H_\phi = -E_0 \frac{j\omega\varepsilon}{h_{nm}} J'_n(h_{nm}\rho)\cos n\phi \left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.53e}$$

where the modal eigenvalues are $h_{nm} = p_{nm}/a$, where the values of p_{nm} are given by Eq. (7.45). Here, E_ϕ must vanish at $z = \pm d/2$, so it is required that $A^- = A^+$ (even modes) or $A^- = -A^+$ (odd modes) and that β_{nm} be restricted to the values $l\pi/d$, where $l = 0, 1, \ldots$. The eigenvalues of the TM_{nm} modes are different than the corresponding TE modes, so the resonant frequencies of the TM_{nml} modes are also different:

$$f_{nml} = \frac{1}{2\pi\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{p_{nm}}{a}\right)^2 + \left(\frac{l\pi}{d}\right)^2} \qquad (TM_{nml} \text{ modes}) \tag{7.54}$$

Figure 7.14 is a resonant mode chart for a cylindrical cavity, which shows the resonant frequencies of the lowest order modes as a function of the cylinder radius to length ratio. Here, it is seen that the TE_{111} mode has the lowest resonant frequency when $a/d < 2$, whereas the TM_{010} mode has the lowest resonant frequency when $a/d > 2$.

An important characteristic of a resonant mode is its quality factor Q, defined as

$$Q = 2\pi f \times \frac{\text{average energy stored}}{\text{power loss}} \tag{7.55}$$

At resonance, the average energies stored in the electric and magnetic fields are equal, so Q can be expressed as

$$Q = \frac{4\pi f_o W_e}{P_L} \tag{7.56}$$

© 2006 by Taylor & Francis Group, LLC

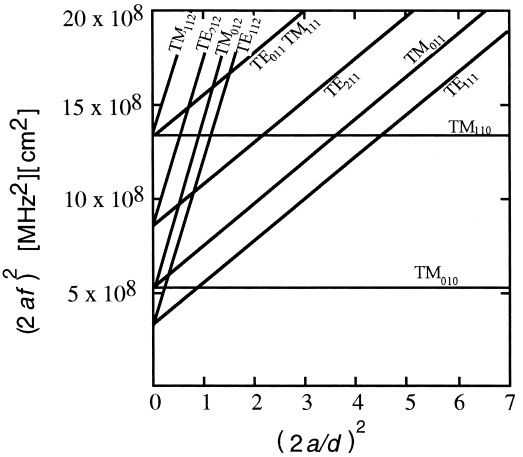


Figure 7.14 Resonant mode chart for cylindrical cavities. (Adapted from Collin, R. *Foundations for Microwave Engineering*; McGraw-Hill, Inc.: New York, 1992, with permission.)

where W_e is the time-average energy stored in electric field and P_L is the time-average dissipated power at resonance. This is the same definition for the quality factor as is used for lumped-element tuned circuits [8]. Also as in lumped circuits, the quality factor Q and the 3-dB bandwidth (BW) of a cavity resonator are related by

$$\mathrm{BW} = \frac{2\pi f_o}{Q} \quad [\mathrm{Hz}] \tag{7.57}$$

where f_o is the resonant frequency of the cavity.

The losses in metal resonators are nearly always dominated by the conduction losses in the cylinder walls. Similar to the way in which waveguide losses are evaluated, this power loss can be evaluated by integrating the tangential H fields over the outer surface of the cavity:

$$\begin{aligned} P_L &= \frac{R_s}{2}\oint_s H_{\tan}^2\, ds \\ &= \frac{R_s}{2}\left\{\int_0^{2\pi}\int_0^{d}\left[|H_\phi(\rho=a)|^2 + |H_z(\rho=a)|^2\right] a\, d\phi\, dz \right. \\ &\quad \left. + 2\int_0^{a}\int_0^{2\pi}\left[|H_\rho(z=0)|^2 + |H_\phi(z=0)|^2\right]\rho\, d\rho\, d\phi\right\} \end{aligned} \tag{7.58}$$

where R_s is the surface resistance of the conducting walls and the factor 2 in the second integral occurs because the losses on the upper and lower end caps are identical. Similarly, the energy stored in the electric field is found by integrating the electric energy density throughout the cavity.

$$W_e = \frac{\varepsilon}{4}\int_0^{a}\int_0^{2\pi}\int_{-d/2}^{d/2}\left(|E_\rho^2| + |E_\phi^2| + |E_z^2|\right)\rho\, d\rho\, d\phi\, dz \tag{7.59}$$

© 2006 by Taylor & Francis Group, LLC

Using the properties of Bessel functions, the following expressions can be obtained for TE_{nml} modes [9]:

$$Q\frac{\delta}{\lambda_o}=\frac{\left[1-\left(n/p'_{nm}\right)^2\right]\left[\left(p'_{nm}\right)^2+\left(l\pi a/d\right)^2\right]^{3/2}}{2\pi\left[\left(p'_{nm}\right)^2+(2a/d)\left(l\pi a/d\right)^2+\left(nl\pi a/p'_{nm}d\right)(1-2a/d)\right]} \qquad TE_{nml}\text{ modes} \quad (7.60)$$

where $\delta=1/\sqrt{\pi f\mu\varepsilon}$ is the skin depth of the conducting walls and λ_o is the free-space wavelength. Similarly, for TM_{nml} modes [9],

$$Q\frac{\delta}{\lambda_o}=\begin{cases}\dfrac{p_{nm}}{2\pi(1+2a/d)} & l=0\\[2ex] \dfrac{\left[p_{nm}^2+\left(l\pi a/d\right)^2\right]^{1/2}}{2\pi(1+2a/d)} & l>0\end{cases} \qquad TM_{nml}\text{ modes} \quad (7.61)$$

Figure 7.15 shows the Q values of some of the lowest order modes as a function of the of the cylinder radius-to-length ratio. Here it is seen that the TE_{012} has the highest Q, which makes it useful for applications where a sharp resonance is needed. This mode also has the property that $H_\phi=0$, so there are no axial currents. This means that the cavity endcaps can be made movable for tuning without introducing additional cavity losses.

Coupling between metal resonators and waveguiding structures, such as coaxial cables and waveguides, can be arranged in a variety of ways. Figure 7.16 shows three possibilities. In the case of Fig. 7.16a, a coaxial line is positioned such that the E field of the desired resonator mode is tangential to the center conductor probe. In the case of Fig. 7.16b, the loop formed from the coaxial line is positioned such that the H field of the desired mode is perpendicular to the plane of the loop. For waveguide to resonator coupling, an aperture is typically placed at a position where the H fields of both the cavity and waveguide modes have the same directions. This is shown in Fig. 7.16c.

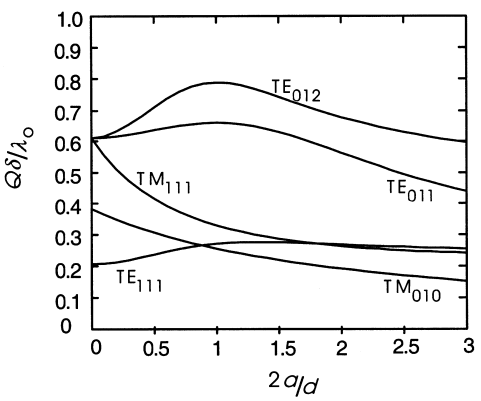


Figure 7.15 Q for cylindrical cavity modes. (Adapted from Collin, R. *Foundations for Microwave Engineering*; McGraw-Hill, Inc.: New York, 1992, with permission.)

© 2006 by Taylor & Francis Group, LLC

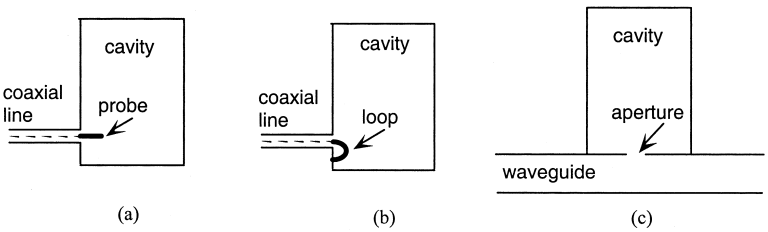


Figure 7.16 Coupling to methods for metal resonators. (a) probe coupling, (b) loop coupling, (c) aperture coupling.

7.9.2. Dielectric Resonators

A resonant cavity can also be constructed using a dielectric cylinder. Like metal cavity resonators, dielectric resonators operate on the principle of constructive interference of multiply reflected waves, but dielectric resonators differ in that some fringing or leakage of the fields occur at the dielectric boundaries. Although this fringing tends to lower the resonator Q values, it has the advantage that it allows easier coupling of energy into and out of these resonators. In addition, the high dielectric constants of these resonators allow them to be made much smaller than air-filled cavity resonators at the same frequencies. A number of dielectric materials are available that have both high dielectric constants, low loss-tangents ($\tan\delta$), and high temperature stability. Typical examples are barium tetratitanate ($\varepsilon_r = 37$, $\tan\delta = 0.0005$) and titania ($\varepsilon_r = 95$, $\tan\delta = 0.001$).

Just as in the case of metal cavity resonators, the modes of dielectric resonators can be considered as waveguide modes that reflect back and forth between the ends of the cylinder. The dielectric constants of dielectric resonators are usually much larger than the host medium (usually air), so the reflections at the air–dielectric boundaries are strong, but have polarities that are opposite to those obtained at dielectric–conductor boundaries. These reflections are much like what would be obtained if a magnetic conductor were present at the dielectric interface. For this reason, the TE modes of dielectric resonators bear similarities to the TM modes of metal cavity resonators, and vice versa.

An exact analysis of the resonant modes of a dielectric resonator can only be performed numerically, due to the difficulty of modeling the leakage fields. Nevertheless, Cohn [10] has developed an approximate technique that yields relatively accurate results with good physical insight. This model is shown in Fig. 7.17. Here, a dielectric cylinder of radius a, height d, and dielectric constant ε_r is surrounded by a perfectly conducting magnetic wall. The magnetic wall forces the tangential H field to vanish at $\rho = a$, which greatly simplifies analysis, but also allows fields to fringe beyond endcap boundaries.

The dielectric resonator mode that is most easily coupled to external circuits (such as a microstrip transmission line) is formed from the sum of upward and downward TE_{01} waves. Inside the dielectric ($|z| < d/2$), these are

$$H_z = H_0 J_o(k_\rho \rho)\left(A^+ e^{-j\beta z} + A^- e^{j\beta z}\right)e^{j\omega t} \tag{7.62a}$$

$$H_\rho = -H_0 \frac{j\beta}{k_\rho} J_o'(k_\rho \rho)\left(A^+ e^{-j\beta z} - A^- e^{j\beta z}\right)e^{j\omega t} \tag{7.62b}$$

$$E_\phi = H_0 \frac{j\omega\mu_o}{k_\rho} J_o'(k_\rho \rho)\left(A^+ e^{-j\beta z} + A^- e^{j\beta z}\right)e^{j\omega t} \tag{7.62c}$$

© 2006 by Taylor & Francis Group, LLC

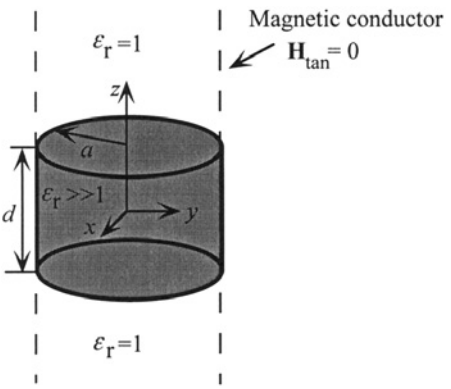


Figure 7.17 Magnetic conductor model of dielectric resonator.

where

$$\beta = \sqrt{\varepsilon_r k_o^2 - k_\rho^2} \tag{7.63}$$

and $k_o = 2\pi f\sqrt{\mu_o \varepsilon_o}$ is the free-space wave number. The value of k_ρ is set by the requirement that H_z vanishes at $\rho = a$, so

$$k_\rho\, a = p_{01} = 2.4048 \tag{7.64}$$

Symmetry conditions demand that either $A^+ = A^-$ (even modes) or $A^+ = -A^-$ (odd modes).

The same field components are present in the air region ($|z| > d/2$), where there are evanescent fields which decay as $e^{-\alpha|z|}$, where the attenuation constant α is given by

$$\alpha = \sqrt{k_\rho^2 - k_o^2} \tag{7.65}$$

Requiring continuity of the transverse electric and magnetic fields at the cylinder endcaps $z = \pm d/2$ yields the following resonance condition [11]:

$$\beta d = 2\tan^{-1}\left(\frac{\alpha}{\beta}\right) + l\pi \tag{7.66}$$

where l is an integer. Using Eqs. (7.63) and (7.65), Eq. (7.66) can be solved numerically for k_o to obtain the resonant frequencies. The lowest order mode (for $l = 0$) exhibits a less-than-unity number of half-wavelength variations along the axial coordinate z. For this reason, this mode is typically designated as the $TE_{01\delta}$ mode.

An even simpler formula, derived empirically from numerical solutions, for the resonant frequency of the $TE_{01\delta}$ mode is [12]

$$f_{\mathrm{GHz}} = \frac{34}{a_{\mathrm{mm}}\sqrt{\varepsilon_r}}\left(\frac{a}{d} + 3.45\right) \tag{7.67}$$

© 2006 by Taylor & Francis Group, LLC

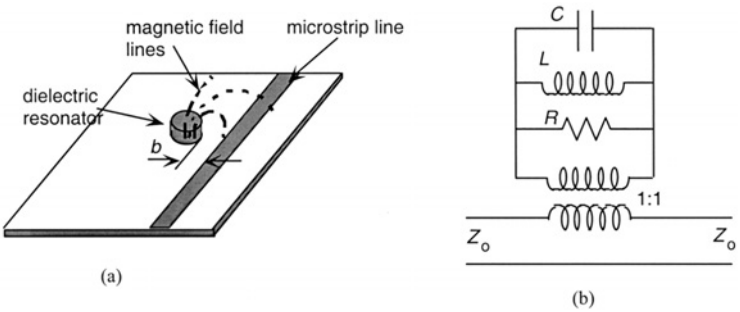


Figure 7.18 (a) Dielectric resonator coupled to a microstrip line and (b) the equivalent circuit.

where a_{mm} is the cylinder radius in millimeters. This formula is accurate to roughly 2% for the range $0.5 < a/d < 2$ and $30 < \varepsilon_r < 50$.

Dielectric resonators typically exhibit high Q values when low-loss dielectrics are used. In that case, radiation loss is the dominant loss mechanism, and typical values for the unloaded Q range from 100 to several thousand. For situations where higher Q values are required, the resonator can be placed in a shielding box. Care should be taken that the distance between the box and the resonator is large enough so that the resonant frequency of the resonator is not significantly affected.

Figure 7.18a shows a dielectric resonator that is coupled to a microstrip transmission line. Here, it is seen that the magnetic fields lines generated by the microstrip line couple strongly to the fringing magnetic field of the $\text{TE}_{01\delta}$ mode. The amount of coupling between the the microstrip line and the resonator is controlled by the offset distance b between the resonator and the line.

The equivalent circuit that the resonator presents to the microstrip line is shown in Fig. 7.18b. In this model, the resonator appears as a parallel resonant circuit, coupled to the microstrip like through a 1:1 transformer. The resonator's resonant frequency f_o and unloaded Q are related to the lumped circuit parameters by the relations

$$f_o = \frac{1}{2\pi\sqrt{LC}} \tag{7.68}$$

$$Q = 2\pi f_o RC \tag{7.69}$$

The effect of the coupling between the resonator and the transmission line is to decrease the circuit Q. The larger the coupling, the smaller the overall Q. The coupling g between the resonator and the transmission line is defined as the ratio of the unloaded Q to the external Q. When both the source and load sides of the transmission line are terminated in matched loads, the external load presented to the resonator is $2Z_o$, so

$$g = \frac{Q}{Q_{\text{ext}}} = \frac{\omega_o RC}{\omega_o(2Z_o)C} = \frac{R}{2Z_o} \tag{7.70}$$

where Z_o is the characteristic impedance of the transmission line. In practice, g can be determined experimentally by measuring the reflection coefficient Γ seen from the source end of the transmission line when both the source and load are matched to

© 2006 by Taylor & Francis Group, LLC

the transmission line. At resonance, the load seen by the source is $Z_o + R$, so the reflection coefficient is:

$$\Gamma = \frac{(Z_o + R) - Z_o}{(Z_o + R) + Z_o} = \frac{g}{1 + g} \tag{7.71}$$

Equations (7.68)–(7.71) can be used to uniquely determine the lumped parameters that a given resonator presents to a transmission line.

REFERENCES

1. Demarest, K.R. *Engineering Electromagnetics*; Prentice Hall: Upper Saddle River, NJ, 1998. Chapter 13, 508–563.
2. Marcuvitz, N. *Waveguide Handbook*, 2nd Ed., Peter Peregrinus Ltd.: London, 1986; 24–25.
3. Collin, R. *Foundations for Microwave Engineering*, 2nd Ed.; McGraw-Hill, Inc.: New York, 1992; 189.
4. Pozar, D. *Microwave Engineering*, 2nd Ed.; Wiley: New York, 2005, 683–686.
5. Abramowitz, M.; Stegun, I. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; Wiley: New York, 1972, 358–374.
6. Kretzschmar, J. Wave propagation in hollow conducting elliptical waveguides. IEEE Transactions on Microwave Theory and Techniques **1970**, *18*: 547–554.
7. Montgomery, J. On the complete eigenvalue solution of ridged waveguide. IEEE Transactions on Microwave Theory and Techniques **1971**, *19*, 457–555.
8. Collin, R. *Foundations for Microwave Engineering*, 2nd Ed.; McGraw-Hill, Inc.: New York, 1992; 325–330.
9. Collin, R. *Foundations for Microwave Engineering*; McGraw-Hill, Inc.: New York, 1992; 506–507.
10. Cohn, R. Microwave bandpass filters containing high-Q dielectric resonators. IEEE Transaction on Microwave Theory and Techniques **1968**, *16*, 218–227.
11. Kajfez, D.; Guillon, P. *Dielectric Resonators*; Artech House, Inc.: Dedham, MA, 1986; 126–132.
12. Kajfez, D.; Guillon, P. *Dielectric Resonators*; Artech House, Inc.: Dedham, MA, 1986; 3.
13. Collin, R. *Foundations for Microwave Engineering*; McGraw-Hill, Inc.: New York, 1992; 508.

FURTHER INFORMATION

There are many textbooks and handbooks that cover the subject of waveguides in great detail. In addition to the references cited above, others include

Baden Fuller, A.J. *Microwaves: An Introduction to Microwave Theory and Techniques*; Pergamon Press: Oxford [England]; New York, 1990.

Cronin, N. *Microwave and Optical Waveguides*; Institute of Physics: Bristol; Philadelphia, 1995.

Liao, S. *Microwave Devices and Circuits*, 3rd Ed.; Prentice Hall: Englewood Cliffs, N.J., 1990.

Pozar, D. *Microwave Engineering*; Wiley: New York, 1998.

Carpentier, M.; Smith, B. *Microwave Engineering Handbook*; Van Nostrand Reinhold: New York, 1993.

© 2006 by Taylor & Francis Group, LLC

Abramowitz, M.; Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; Wiley: New York, 1972.
Lewin, L. *Theory of Waveguides*; Wiley: New York, 1975.
Collin, R.; Field, R.E. *Theory of Guided Waves*; 2nd Ed., IEEE Press: Piscataway, NJ, 1991.
Gardiol, F. *Introduction to Microwaves*; Artech House, Inc.: Dedham, MA, 1984.
Ramo, S.; Whinnery, J.; Van Duzer, J. *Fields and Waves in Communication Electronics*; Wiley: New York, 1994.

© 2006 by Taylor & Francis Group, LLC

8
Antennas: Fundamentals

David Thiel
Griffith University
Nathan, Queensland, Australia

8.1. INTRODUCTION TO RADIATION

Electromagnetic radiation is one of the principal forms of conveying information from one point to another—from person to person, computer to computer, telephone to telephone and broadcast radio station to radio receiver. The radiation used in these communications systems usually lies in the frequency range from extremely low frequencies (ELF) to optical and ultraviolet (UV) frequencies. For example, ELF radiation (frequency band 3 Hz to 3 kHz) is used in through-earth propagation and telephone modems. Optical and UV frequencies are commonly used with optical fibers and sometimes in open-air links. Electromagnetic radiation can be trapped and directed along conductive wires (transmission lines), dielectric filled conducting pipes (wave guides), and in dielectric pipes sheathed with dielectric materials with a lower dielectric constant (optical fibers).

In many cases it is desirable to have a wireless EM link so that the radiation is unguided and will generally follow a line-of-sight path (i.e., a geometrical optics path). In the radio-frequency (RF)–microwave-frequency range, antennas are often used to launch and focus the radiation to a limited beam width so that the signal to noise ratio at the receiver is maximum and the interference to other wireless links in the same frequency band is minimized. An antenna is therefore a device that converts confined radiation from a transmission line or waveguide into an unguided but directed electromagnetic wave in the ambient medium (often, but not always, air).

While electromagnetic waves can propagate along the interface between media (e.g., surface waves) and in waveguides (e.g., TE and TM waves) in such a way that the electric and magnetic fields are not perpendicular to the direction of propagation, in most cases, the free-space radiation is in the form of a transverse electromagnetic (TEM) wave. For example, for a TEM wave, if we choose the electric field vector **E** to lie in the direction of the z axis, the magnetic field vector **H** to lie in the x direction, then the direction of propagation can be defined by the wave vector **k**, which lies in the y direction. This is shown in Fig. 8.1.

The electric field strength **E** has the units of V/m, the magnetic field strength **H** has units A/m, so the power density of the electromagnetic wave is given by the

© 2006 by Taylor & Francis Group, LLC

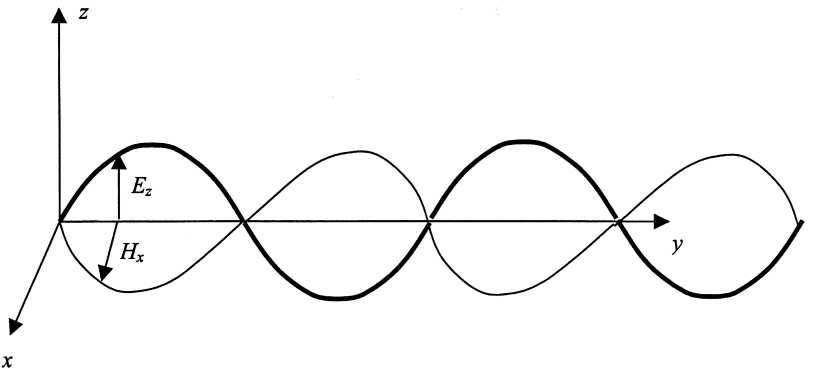


Figure 8.1 TEM wave with axis definition.

Poynting vector **S**, where

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} \tag{8.1}$$

S has the units of watts per square meter and so is a measure of the power density of the radiation.

The radiation is described as being linearly polarized if the direction of the **E** field remains constant along the path of propagation.

In a TEM wave, it is possible for the direction of **E** to vary continuously in the *xz* plane perpendicular to the direction of radiation. In this case the radiation scribes an ellipse or circle as it propagates, and the radiation is called *elliptical* or *circular polarization*. Simple vector addition can show that elliptically polarized radiation has a linearly polarized component and a linearly polarized wave has a circularly polarized component.

The electric field vector **E** defines the force exerted on a charged particle in the presence of a TEM wave. If the charged particles are free to move, e.g., as electrons on the surface of a good conductor, a current is induced on the wire, and this can be detected and processed using standard electronic circuit techniques. Clearly, if a linearly polarized TEM wave has an **E** component in the *z* direction, then a straight wire in the *z* direction will have maximum current induced, whereas a wire directed in the *x* or *y* direction will have zero current. Thus a simple straight wire can be used to detect the presence on an electromagnetic wave, and so a wire is the basic form of a linearly polarized antenna.

The receiving characteristics of an antenna are identical to its transmitting characteristics; thus, descriptions of the properties of an antenna are equally valid in terms of the reception characteristics and transmission characteristics. This property is described in terms of the reciprocity principle for a communications link in which the transmitting and receiving antennas can be exchanged and the signal strength into the receiver is unchanged providing there is no media boundary in the vicinity of the antennas.

8.2. ANTENNA TERMINOLOGY

There are many different requirements for antenna systems. In broadcast applications (e.g., radio, television), it is desirable that the transmitted radiation can be detected over a large area. In point to point applications (e.g., fixed microwave link, communications with a fixed earth station and a geostationary satellite), it is desirable to confine the transmitted

© 2006 by Taylor & Francis Group, LLC

radiation to a small angle. In mobile communications applications, a point-to-point communications link is required, but the location of one point can move continuously during the transmission. In the case of multiple in–multiple out systems (MIMO), the location of the antennas can be quite varied. Recently there is increased interest in ad hoc radio networks or unplanned networks that self-assemble using smart antennas.

In designing a communications system, it is necessary to calculate the radiation strength at the receiver to ensure adequate signal-to-noise ratio for the correct interpretation of the signals received. This is called a *link budget calculation*. Therefore, it is necessary to specify the directional characteristics of the antenna in such a way that the power received by the target receiver can be calculated from the power delivered to the input terminals of the transmitting antenna.

It is convenient to specify the principal radiation direction in terms of a spherical polar coordinate system centered on the transmitting antenna. In Fig. 8.2, the principal radiation direction (main beam direction) of the antenna is (ϕ_0,θ_0), and the strength of the radiation in other directions is plotted as a three-dimensional surface. In this polar plot, the distance from the origin of the coordinate system (the phase center of the antenna) and a point on the surface represents the radiation field strength in that particular direction when measured in the far field, i.e., some considerable distance from the antenna. The spread in the radiation field can be defined in terms of the angular displacement from the principal direction of radiation where the field strength falls to one half the power (or -3 dB) in the principal direction. These half power points define the two beam widths $\Delta\theta$ and $\Delta\phi$ as shown in Fig. 8.2.

The directional characteristics can be described by two-dimensional and three-dimensional radiation patterns, in which the relative signal strength is plotted as a function of angle. The field strength is usually plotted in dBi. This is the gain in dB relative to an isotropic radiator having the same power output. As the radiation pattern represents the three-dimensional gain as a function of two angular directions (θ and ϕ), it is common to define a plane (e.g., the $\theta = 90^\circ$ plane) and plot the signal strength for all angular positions

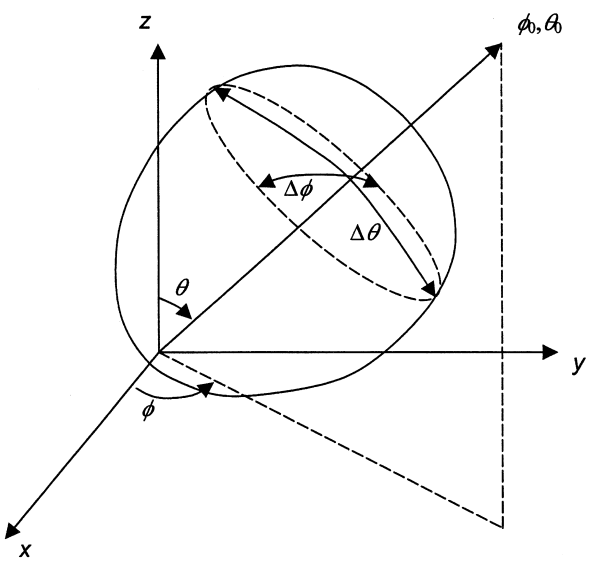


Figure 8.2 Main beam direction definition and beam width.

© 2006 by Taylor & Francis Group, LLC

in that plane (e.g., all values of ϕ for the $\theta = 90^\circ$ case). Usually that plane includes the origin of the coordinate system or the phase center of the antenna. In some applications such as an antenna located just above an infinite ground plane, the cut-plane for the radiation pattern includes the maximum radiation gain, which is elevated from the ground plane. In this case the radiation pattern is taken at a fixed elevation angle above the phase center of the antenna. In Fig. 8.2, this elevated radiation pattern would be located at $\theta = \theta_0$ for all values of ϕ.

Figure 8.3 is an example of a two-dimensional radiation pattern in which the -3 dB beam width is defined. The front-to-back ratio FB of an antenna is another important characteristic and is defined in terms of the ratio of the field strength in the direction (θ_0,ϕ_0) to the field strength in the opposite direction $(180^\circ - \theta_0,\phi_0 + 180^\circ)$. FB is usually defined in dB.

The directivity of an antenna is the ratio of the power density in the main beam to the average power density (i.e., total radiated power divided by 4π) [1]. The larger the value of the directivity, the more directional is the antenna. The directivity is always greater than 1.

The antenna efficiency is similar to the directivity but also includes losses in the antenna structure (e.g., the effect of finite conductivity, dielectric losses and sometimes even the impedance mismatch with the transmission feed line).

Antenna gain is the ratio of the radiation intensity in the main beam to the radiation intensity in every direction assuming that all radiated power is evenly distributed in all possible directions [1].

An antenna can also be described in terms of a circuit element connected to a transmission line. The input impedance of an antenna Z_a has both real and imaginary parts—the real part R_a relates to the loss of energy due to the radiated field and material losses; the imaginary part X_a relates to the inductive or capacitive load that the antenna structure presents. Maximum power transfer is achieved when the antenna input impedance is equal to the characteristic impedance Z_0 of the transmission feed line across the frequency range of interest. That is,

$$Z_a = R_a + jX_a = Z_0 \tag{8.2}$$

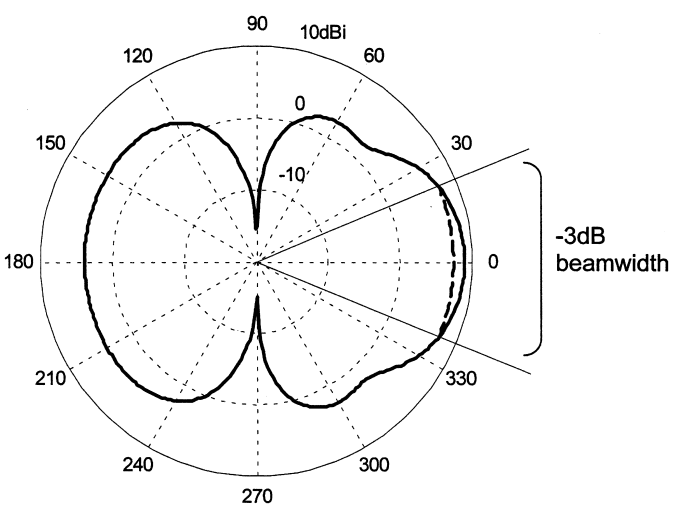


Figure 8.3 Typical two-dimensional radiation pattern illustrating the 3-dB beamwidth.

© 2006 by Taylor & Francis Group, LLC

If there is an impedance mismatch, then there is reflection of the signal back into the transmission line. This is commonly described in terms of the scattering parameter S_{11} and can be determined using Eq. (8.3)

$$S_{11} = 20\log_{10}\left|\frac{Z_a - Z_0}{Z_a + Z_0}\right| \tag{8.3}$$

The resonant frequency f_0 of the antenna can be defined as that frequency where the reactance of the antenna is zero [1]. This can be shown to be true when the S_{11} value is a minimum. The frequency bandwidth of an antenna is commonly defined as the frequency range where the S_{11} value is less than -10 dB. In numerical terms, this definition implies that the antenna constitutes an impedance mismatch that reflects less than 10% of the power back into the transmission line.

It is possible and sometimes desirable to define the resonant frequency of an antenna in terms of the radiation pattern or antenna gain rather than the impedance. This approach allows the designer to devote most attention to the radiation characteristics of the antenna rather than the impedance matching. It is possible to construct impedance matching circuits to reduce the impact of S_{11} on the link budget. Quarter wavelength chokes are one matching technique used with coaxial cables and microstrip lines [2,3].

The power delivered to the transmission line connected to the feed point of a receiving antenna has been extracted from the radiation falling on the antenna. The radiated field strength is measured in watts per square meter so that one can define the effective area of an antenna illuminated by the incoming radiation, even when the physical area of the antenna structure is very small. In some cases, such as a parabolic dish antenna or other aperture antennas, the antenna area is obvious. For wire antenna structures it is not so obvious, and the effective antenna area must be calculated from the antenna gain assuming uniform radiation is incident over the surface of the antenna [1,2].

Note that in receiving the power from an incoming radio wave, currents are excited in the antenna, which, in turn, cause the receiving antenna to radiate. Thus the maximum energy harvested from an antenna is one-half the energy falling on the antenna.

8.3. SIMPLE ANTENNA STRUCTURES

From the principle of reciprocity, it is possible to describe antennas in terms of their transmission or reception characteristics. In this section we will focus on reception characteristics—that is, the conversion of an incoming TEM wave to a current on a transmission line.

A linearly polarized TEM wave with an electric field component parallel to a conducting wire will induce a current to flow in the wire. This current is maximized if the wire forms part of a resonant circuit at the frequency of the incoming radiation. Thus a straight wire in air having length $l = \lambda/2$, with a transmission line connected to its center point has a fundamental resonance frequency f given by the equation

$$f = \frac{nc}{2l} \tag{8.4}$$

where $n = 1$. There are additional resonant frequencies for positive integer values of n.

© 2006 by Taylor & Francis Group, LLC

At the resonant frequency, the current in the antenna is a standing wave. The RMS current along the length of the antenna element is one-half of a sinusoid with maximum current in the center and zero current on the ends. The voltage distribution on the antenna is approximately one half cosinusoidal so the impedance of the antenna at the center feed point is maximized. As noted before, there is maximum power transfer from the antenna to the transmission line when the antenna impedance is identical to the characteristic impedance of the transmission line. This can be achieved by using standard transmission line matching techniques, by adjusting the feed position along the wire, or by adjusting the inductive load on the antenna by changing the wire thickness or the wire length.

A short thin wire has a cosinusoidal radiation pattern in the plane containing the antenna wires (see Fig. 8.4). This is referred to as the *E-plane radiation pattern* as it lies parallel to the **E** field vector of the radiation from the antenna. There is no preferred direction in the plane perpendicular to the antenna wire, the so-called *H-plane radiation pattern* as the structure is completely symmetric about this line. The input resistance R_a for an electrically short dipole wire (i.e., $l \ll \lambda/2$) is given by [2,3]

$$R_a = 20\pi^2 \left(\frac{l}{\lambda}\right)^2 \tag{8.5}$$

The corresponding radiation pattern is found in all major antenna textbooks [2–5]. Also, the antenna impedance is linearly related to the length of the element provided the inequality

$$l < \frac{\lambda}{3}$$

remains valid.

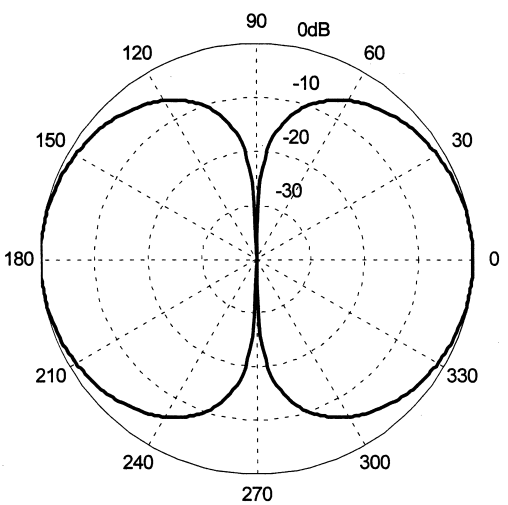


Figure 8.4 E-plane radiation pattern of a Hertzian dipole in dB. The gain has been normalized to 0 dB.

© 2006 by Taylor & Francis Group, LLC

For a resonant straight-wire antenna, $l = \lambda/2$, the radiation pattern is still dependent on θ only and is given by

$$E(\theta, \phi) = \frac{\cos(\sin(\theta)\pi/2))}{\sin\theta} \tag{8.6}$$

and the antenna impedance

$$Z = 73 + 42.5j\,\Omega$$

This antenna is called a *half-wave dipole* and its radiation pattern is shown in Fig. 8.4 [2–5]. If the length of the antenna is reduced slightly, the imaginary part of the impedance can be reduced to zero and the antenna resonates with an input impedance which has a real component only [3].

A long straight wire with length l has a radiation pattern given by

$$E(\theta) = E_0 \frac{\cos(\cos(\theta)kl/2) - \cos(kl/2)}{\sin\theta} \tag{8.7}$$

where the wave number $k = 2\pi/\lambda$. Note that when the size of a radiating structure exceeds λ in one or more dimensions, the radiation pattern has side lobes and nulls. This is illustrated in Fig. 8.5 for a number of center-fed thin-wire antennas with different lengths.

A wire antenna located in the vicinity of a ground plane has its radiation pattern and impedance influenced by the ground plane because currents are induced to flow in the conductor. The simplest approach to understanding this type of antenna structure is to imagine that the ground plane can be replaced by an image antenna element which is located equidistant below the plane. This is illustrated in Fig. 8.6. The vertical current components are in phase with the source currents and the horizontal current components

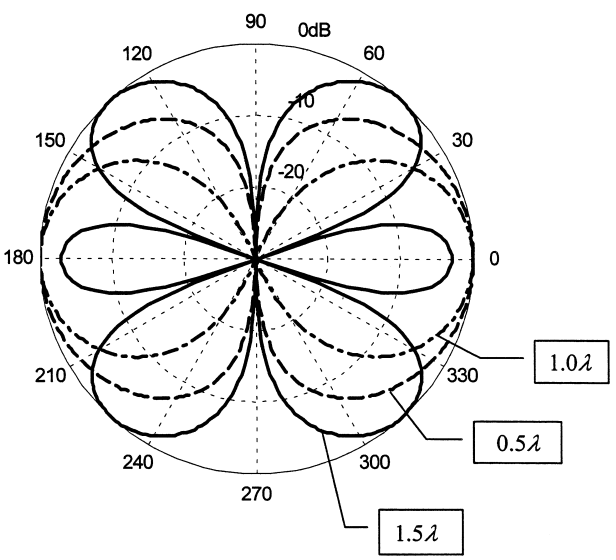


Figure 8.5 Radiation patterns for a number of thin-wire dipole antennas. The antennas lengths are 0.5 λ, 1.0 λ, and 1.5 λ as shown. All gains have been independently normalized to 0 dB.

© 2006 by Taylor & Francis Group, LLC

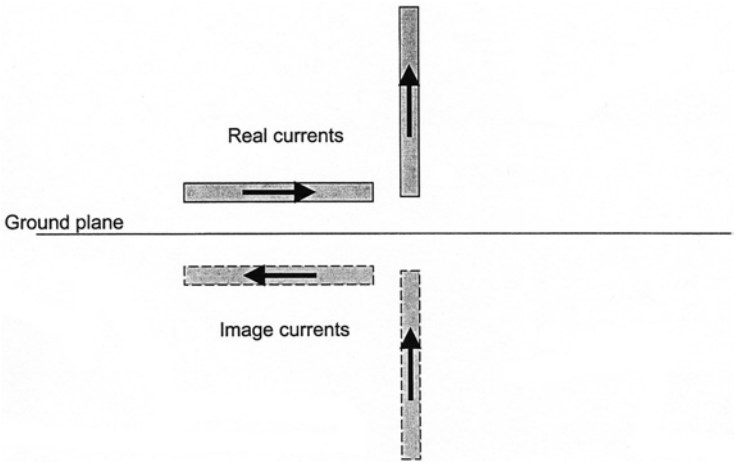


Figure 8.6 Current image elements reflected in the perfectly conducting ground plane of infinite extent. Note that currents normal to the ground plane have an image current in the same direction whereas horizontal currents have an image current in the opposite direction.

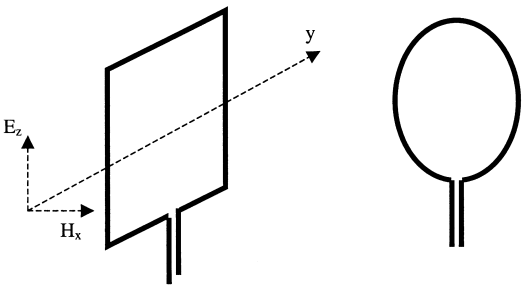


Figure 8.7 Simple loop antenna structures with balanced transmission lines.

are 180 degrees out of phase with the driven element. Thus a vertical wire element of length $\lambda/4$ with one end located on the ground plane has the radiation pattern of a half wave dipole in the hemisphere above the plane. The input impedance of this element is one-half that of the half-wave dipole. This antenna configuration is referred to as a quarter-wave monopole [2–5].

An alternative approach to constructing radiating structures is to use a conducting loop of wire. This can be considered to react to the magnetic field component of a TEM wave. The **H** field component of the radiation drives currents to circulate in the loop. Two simple, single turn, loop antenna structures are illustrated in Fig. 8.7. The current induced in a loop antenna can be increased by increasing the number of turns of wire in the loop structure. When the circumference p of the loop is very much smaller than λ, one obtains maximum response (i.e., the principal radiation direction) when **H** of the TEM wave is perpendicular to the plane of the loop (the yz plane in Fig. 8.7). This antenna is linearly polarized.

Larger sized loops with $p > \lambda$, are mainly used as folded dipole structures (see Fig. 8.8), which have the directionality of the equivalent dipole antenna but with a modified input impedance [2,3].

© 2006 by Taylor & Francis Group, LLC

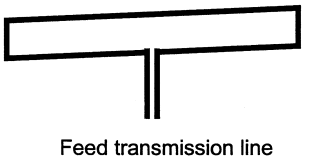


Figure 8.8 Folded dipole loop antenna.

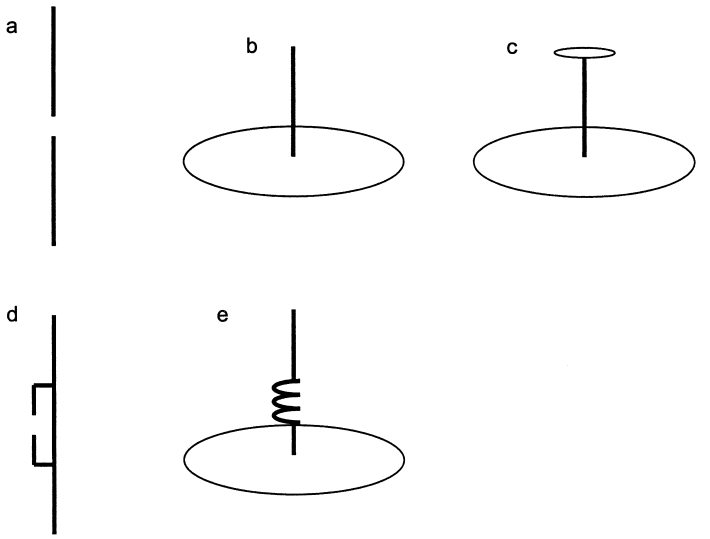


Figure 8.9 Variations on a straight wire antenna. (a) center-fed dipole antenna, (b) end-fed monopole antenna above a ground plane, (c) capacitive loaded monopole antennas, (d) T-feed dipole antennas, and (e) coil-loaded monopole antenna.

All other wire antennas can be described as variations of one of these basic antenna types: the wire monopole on a ground plane, a wire dipole in free space, a wire loop in free space, or a half loop attached to a ground plane. For example, Fig. 8.9 illustrates a number of variations of a straight wire antenna. Figure 8.9a is a center-fed straight thin-wire dipole. Figure 8.9b is a straight thin-wire monopole located on a perfectly conducting ground plane. The feed point of the antenna is at the base of the straight wire. Figure 8.9c is a capacitively loaded thin wire antenna (sometimes called a *capacitive plate antenna* [2]). The conductive disk at the top of the antenna is used to alter the input impedance of the antenna [2]. Figure 8.9d is a T-match configuration for a dipole antenna [2]. In this case the input impedance seen by the transmission line is modified by the feed position on the main antenna element. Figure 8.9e illustrates an end-fed monopole antenna on a ground plane which has a wire coil located partway along its length. This coil has the effect of providing a delay line between the lower straight wire element (usually $\lambda/4$) and the upper straight wire element (usually $\lambda/2$) to provide more gain in the horizontal direction. There are many other variations to straight wire, end-fed, monopole antennas on ground planes and center-fed dipole antennas. The gains of these antennas can be further improved through the use of reflecting planes, corner reflectors, and parasitic elements [2–5].

© 2006 by Taylor & Francis Group, LLC

In order to respond to circularly polarized radiation, two half-wave dipoles oriented perpendicular to each other and perpendicular to the direction of the radiation will have maximum response to circular polarized radiation when one is fed 90 degrees out of phase with the other. When the phase shifter is deployed in the feed line of the other dipole, the sense of the circular polarization is reversed; i.e., right-hand circular polarization becomes left-hand circular polarization. Figure 8.10a shows a simple planar spiral antenna in which the arms of a dipole antenna have been shaped to respond to circularly polarized radiation. This antenna is fed by a balanced transmission line at the two terminals in the center of the antenna. The geometry of this antenna corresponds to the straight-line approximation to an Archimedean spiral [3]. The principal radiation direction is out of the plane. Figure 8.10b shows a helical antenna on a ground plane designed to respond to circularly polarized radiation [2]. The geometry of the helix (i.e., the radius, the number of turns, the total length of the wire, the diameter of the wire, and the pitch of the spiral) has a significant effect on the directional characteristics of the antenna [2,3].

One can increase the effective length of a wire antenna by embedding it in dielectric material (see Fig. 8.11a). A thin coating of dielectric on an end-fed monopole element results in the launching of a trapped surface wave mode in the dielectric. The effective wavelength for this trapped mode $\lambda_g < \lambda$, and the length of the resonant antenna is effectively reduced. The resonance condition required that the length of the monopole is approximately $\lambda_g/4$. The size reduction is dependent on the relative permittivity of the dielectric and the thickness of the coating [5].

In Fig. 8.9c, a top-loaded monopole is illustrated. If the top plate is sufficiently large compared to λ, then a waveguide mode is set up between the top plate and the ground plane. When this propagating wave reaches the end of the parallel plate waveguide, i.e., the edge of the top plate, energy is reflected back toward the source, and a standing wave can be set up between the two plates. Some energy, however, leaks past this termination and is launched as a linearly polarized TEM wave normal to the plane of the patch. This is the basis of a patch antenna element [2–5]. For a simple linearly polarized patch antenna, one can image that the two ends of the patch where the current is zero, are effectively two parallel radiating slots. As the two slots radiate in phase at the resonant frequency, the calculation of the radiation pattern is based on double slit interference where the separation distance between the two slits is approximately the length of the patch.

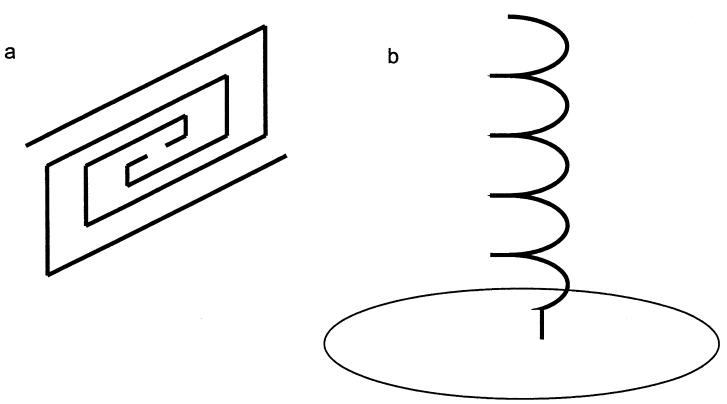


Figure 8.10 Circularly polarized wire antennas. (a) center-fed planar spiral antenna and (b) end-fed helical antenna above a ground plane.

© 2006 by Taylor & Francis Group, LLC

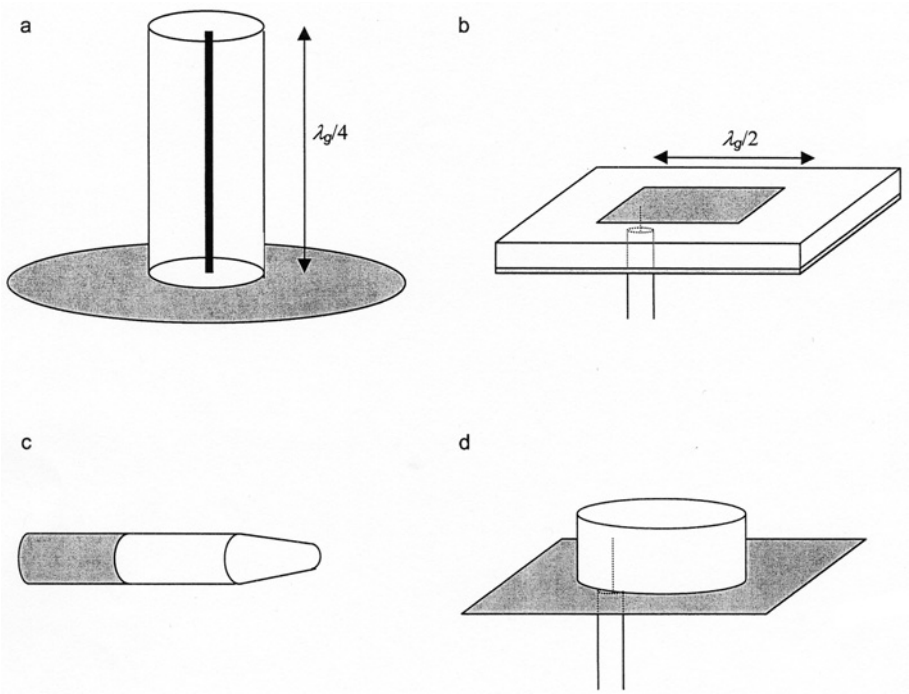


Figure 8.11 Antenna structures incorporating dielectric materials: (a) embedded monopole antenna on a ground plane, (b) patch antennas with coaxial feed probe, (c) dielectric rod antenna as the end of a circular waveguide, and (d) cylindrical dielectric resonator antenna on a ground plane with a coaxial feed probe.

These two edges of the patch are referred to as the *radiating edges*. The other two edges present no significant discontinuity to the current and so do not radiate. Note that with the probe fed patch antenna illustrated in Fig. 8.11b, there are two possible current directions. If the patch is square, then the position of the feed will determine which horizontal direction will provide the best impedance match to the transmission feed line. If the feed probe is located in the center of the patch in one direction, then the current will be maximized for this direction and the antenna impedance at this point is very small. A feed position offset from the center point along one axis can be chosen to provide a near perfect match to a standard coaxial transmission line.

If the patch is rectangular, then there are two possible resonant conditions at different frequencies. The radiation at the two frequencies will be linearly polarized but in orthogonal directions. The effectiveness of the antenna depends on the position of the feed point and the S_{11} of the impedance match.

Patch antenna structures (i.e., a single patch or multiple patches) can be manufactured using standard printed-circuit board photolithographic techniques. The relative permittivity of the substrate material controls the wavelength in the parallel plate waveguide. If the length of the patch is equal to $\lambda_g/2$ where λ_g is the wavelength of the radiation in the waveguide, the patch resonates and the launching efficiency of the antenna is high. The thickness and the relative permittivity of the substrate both have a significant effect on the bandwidth of the antenna. If the substrate is too thick, then the radiation efficiency is

© 2006 by Taylor & Francis Group, LLC

low because a trapped surface-wave mode can propagate through the substrate even when there is no metalization on the upper surface [2–5].

It is possible to launch a circularly polarized TEM wave from a square patch antenna through the use of two orthogonal feed points to create four radiating edges. Alternatively it is possible to use a single feed point by altering the overall shape of the patch with internal slots, truncated corners, etc., to make the propagation in the wave guide circular. Alternatively triangular, circular or other patch shapes can be used. Patch antennas can also be driven using microstrip lines and aperture coupled resonating elements.

A resonant patch can be achieved using a $\lambda_g/4$-long patch terminated along one radiating end with a short-circuit plane connected directly to the ground. In this case, it is important that the current is shorted along the entire length of the patch. This is a slightly more complex problem when constructing these patches.

Another class of antenna uses shaped dielectrics to either guide the radiation as a propagating wave along an elongated structure (a traveling wave dielectric antenna) or as a leakage of radiation from a resonating, compact dielectric structure (see Fig. 8.11c and d, respectively). In the first case, radiation from a waveguide is directed into the shaped dielectric rod where it escapes. The length and the shape of the dielectric material determine the characteristics of the radiation. Commonly, the dielectric material is several wavelengths long.

Figure 8.11d is referred to as a *dielectric resonator antenna* (DRA) where a probe feed is directed through a ground plane into a dielectric cylinder, hemisphere, or cube. The high-permittivity material acts as a resonator with some radiation leaking from the surfaces. The higher the relative permittivity of the dielectric material, the narrower is the impedance bandwidth of the DRA and the smaller is the structure.

The radiation characteristics of all antenna structures can be modified by the close proximity of undriven conducting and dielectric materials. A conducting resonant wire ($l = \lambda/2$) located close to a radiating half-wave dipole acquires an induced current, which means that the element (termed a *parasitic element*) contributes to the impedance of the antenna and its radiation pattern [5]. This effect is termed *mutual coupling*. Effective parasitic elements can be constructed from wires, patches and slots providing they are close to resonance. It is possible to design single feed, multiband antennas with two or more parasitic elements that resonate at frequencies different to the driven element. It is also common to use parasitic elements with resonance points close to that of the center frequency. This allows the total antenna to radiate with an enhanced impedance bandwidth. A Yagi-Uda antenna [2–5] in its traditional form made from wire antenna elements, consists of a center-fed half-wave dipole with a slightly longer "reflector" parasitic element behind it and a number of slightly shorter "director" parasitic elements in the direction of propagation (see Fig. 8.12). This type of antenna has increased bandwidth and narrower beamwidth when compared to a half-wave dipole in isolation.

Planar antennas can be fabricated using printed-circuit board photolithographic processes, which greatly reduces the cost of fabricating large arrays in addition to providing antennas which are conformal with the surface of the support structure. Conformal antennas are often desirable for aesthetic reasons and can be mounted discretely presenting minimal wind resistance. They can also be sealed from the environment using a waterproof coating. If the relative permittivity of the coating is greater than one, then the coating will decrease the resonant frequency of the antenna.

© 2006 by Taylor & Francis Group, LLC

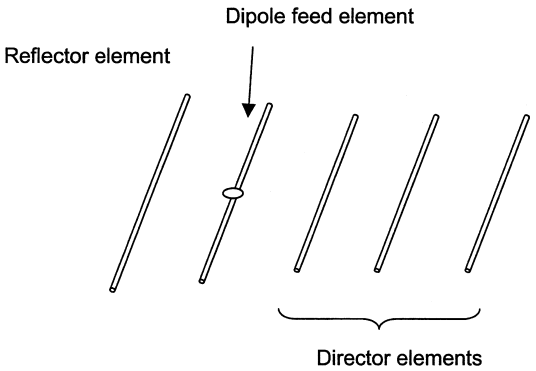


Figure 8.12 Five-element Yagi Uda antenna.

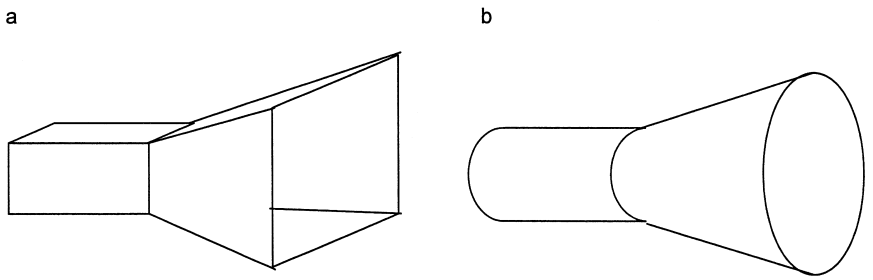


Figure 8.13 Aperture antennas connected to waveguides: (a) pyramidal horn and (b) circular horn antenna.

Another basic form of radiating system is an aperture antennas. These antennas are fed with a waveguide, and the antenna unit consists of a fitting at the end of the waveguide for impedance matching and directivity. Commonly the antenna is a rigid metallic fitting which increases the aperture size from that of the original waveguide. The flare can be in the form of a pyramidal horn or a circular horn (Fig. 8.13a and b, respectively). To reduce the front-to-back ratio, the inside surface of the horn can be treated with slots or dielectric coating so that the currents flowing on the inside of the horn do not flow on the outside surface. The presence of current on the outer side of the horn can greatly increase the back lobe and so decrease the front-to-back ratio of the antenna.

The final category of antenna we will discuss is reflector antennas. In this case, any of the antennas described previously may be placed at the focus of a conducting (and so reflecting) parabolic section. The radiation from such a combination is predominantly a parallel beam with the side-lobe levels and the nulls being principally determined by the size of the reflector. If the feed antenna is located along the axis of the parabolic surface it will lie in the path of the radiation. This effect is referred to as *feed blockage*, and the gain of the antenna is decreased. For this reason, the feed antenna at the focus should be as small as possible or offset from the main beam of the antenna.

All antennas have a radiation pattern that can be described by a function $F(\theta,\phi)$ and input impedance Z_a. It is possible to improve the directivity and gain of the antenna by using a number of identical elements all oriented in the same direction. Such arrangements

© 2006 by Taylor & Francis Group, LLC

are referred to as *antenna arrays*, and a number of common array configurations are discussed in the next section.

8.4. ANTENNA ARRAYS AND PATTERN SYNTHESIS

The gain of an antenna system can be increased significantly if many antennas are positioned in a simple, regular, geometrical configuration such as a straight line, a circle, or a plane (Fig. 8.14). By varying the amplitude and phase of the currents I_n, in the

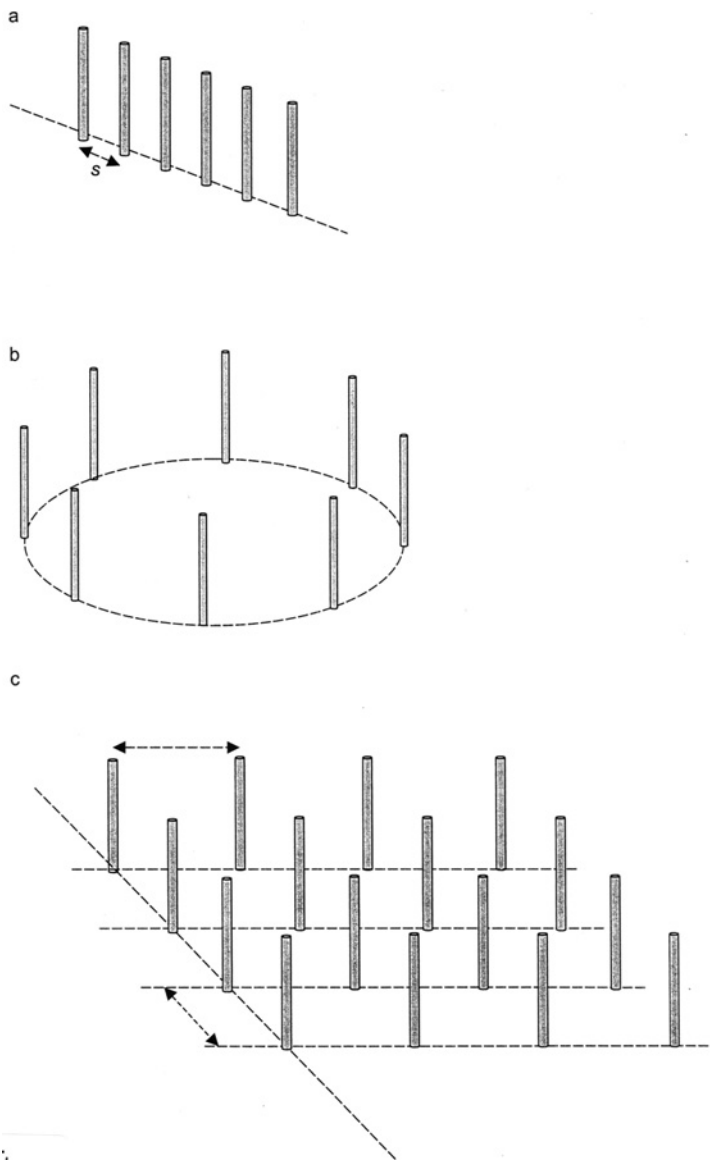


Figure 8.14 Regular monopole array structures: (a) linear array, (b) circular array, and (c) regular planar array.

© 2006 by Taylor & Francis Group, LLC

elements of the array, the direction of the radiation can be altered. As an illustration, consider an equally spaced (along the z axis) linear array of N elements (Fig. 8.14a). Each element in the array is fed with the same current magnitude I with a regular stepped increase in phase of the elements ξ. Assume that each element has directional characteristics $F(\theta,\phi)$ and the antennas are sufficiently far apart ($s > \lambda$) to ensure that mutual coupling between adjacent elements is sufficiently small to ensure they radiate independently.

The radiation pattern $E_{tot}(\theta,\phi)$ for the array is the sum of the phase shifted fields from each element. Thus we can write

$$E_{tot}(\theta,\phi) = IF(\theta,\phi) + (\theta,\phi)e^{jks\cos\theta+\xi} + IF(\theta,\phi)e^{j2ks\sin\theta+2\xi} + \cdots IF(\theta,\phi)e^{jNks\cos\theta+N\xi}$$

$$= F(\theta,\phi)\frac{\sin N\varphi}{N\sin\varphi} = F(\theta,\phi)F_A(\theta,\phi)I \tag{8.8}$$

where $\varphi = ks\cos\theta + \xi$.

The radiation pattern of the array is the product of the element radiation pattern $F(\theta,\phi)$ and the array factor $F_A(\theta,\phi)$ while the radiation intensity is directly proportional to the current magnitude I.

While F_A may be nonzero at a particular angular location (θ,ϕ), if this is the position of a null in the element factor $F(\theta,\phi)$, then clearly there will be a null in $E_{tot}(\theta,\phi)$ at this same angular position.

In extreme cases, the element spacing s can be a significant portion of the circumference of the earth. This type of array has a very narrow beam width and is referred to as an *interferometer array*. The most difficult task in these large arrays is ensuring phase coherence between the spaced receivers.

From the array analysis given above, a change in ξ will change the θ direction of the array. This in turn will change the position of the nulls and usually changes the beam width of the main beam. In a two-dimensional array of antennas, the direction of the main beam can be changed in both θ and ϕ directions using appropriate phase shifts. Careful control over the amplitude and the phase of the current in each element individually can improve the side-lobe levels and the gain of the antenna system. This requires significant computation if the number of elements in the array is large.

Note that when the elements in the array are too close, i.e., s is too small, then mutual coupling can restrict the size of the phase difference between adjacent elements regardless of the feed voltages. This can result in unexpected nulls in the radiation pattern. This effect is referred to as *scan blindness*.

The analysis used to derive the directional characteristics of an array can also be used in reverse; that is, if the radiation pattern is known, it is possible to calculate the magnitude and phase of the currents required on each element to generate such a pattern [2]. An alternative view of this process is to recognize that the far field radiation pattern is given by the two-dimensional Fourier transform of the spatial distribution of the antenna element currents. The current distribution in the array is the inverse Fourier transform of the far field radiation pattern.

Unfortunately this antenna design process is not simple as the far field pattern is of infinite extent and the phase distribution of the fields is unknown. This means an exact solution cannot be obtained and an iterative optimization procedure is required to obtain the best solution based on a defined cost function [5].

© 2006 by Taylor & Francis Group, LLC

This optimization technique has been used to arrange appropriate satellite coverage of the more densely populated areas in the continents of the world. The array of waveguide manifolds and horns can be quite complex involving a large number of apertures with appropriate phase shifting waveguide lengths and power splitters and combiners.

8.5. SMART ANTENNAS

Electronically controlled antennas have been described as smart antennas. In such a system, the main beam direction, resonant frequency (or frequencies for multiband antennas), and null positions are altered electronically to ensure optimum signal-to-noise/interference (SNI) ratio. It is possible to achieve this control through the use of programmable attenuators and phase shifters in phased arrays or the feed position and resonant status of parasitic elements in switched parasitic antennas [5].

In the section on phased arrays, it was demonstrated that the main beam position is controlled by the magnitude and phase of the current in each element. Programmable phase shifters can be constructed using p.i.n. diodes or micro-electro-mechanical radio frequency (MEM RF) switches [6,7] to change the electrical length of the feed transmission line or by applying voltages to electrically active ferrite materials in waveguides. Note that these technologies have some level of insertion loss and the precise phase shift is difficult to control during manufacture. This usually means that the system, once constructed, must be calibrated precisely. Figure 8.15 shows the layout of a serpentine digital phase shifter based on open- and short-circuited switches. If N is the number of bits in the phase shifter, then there are 2^N switch positions, and the lengths of the transmission line elements are

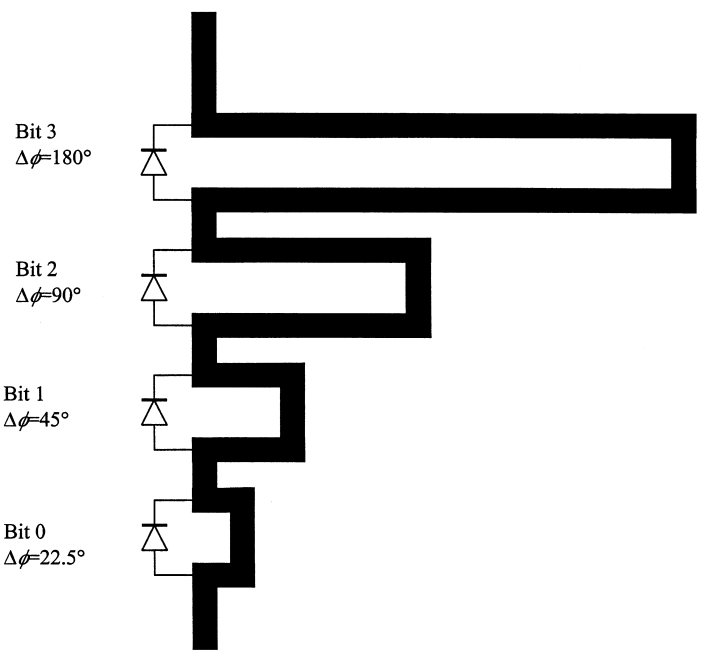


Figure 8.15 Serpentine digital phase shifter electrically controlled by p.i.n. diodes that short-circuit the microstrip line with an applied voltage.

© 2006 by Taylor & Francis Group, LLC

$360n/N$ degrees, where $n = 1, 2, \ldots, N-1$. A change in the least significant binary bit corresponds to a $360/N$-degree phase shift. A similar technique can be used to fabricate a digital attenuator in which different resistive loads are switched in and out of the transmission line.

Switched parasitic antenna principles can be illustrated simply using a 2- and 3-dipole antenna system [5]. In the switched active, switched parasitic antenna (SASPA), an RF switch is used to change the position of the feed from one dipole to the other (see Figure 8.16a). The inactive element is short circuited at its center to provide a reflector element. This is shown as F/S in Fig. 8.16a. In this way the main beam direction is changed by 180 degrees (Fig. 8.17). Note that the F/S combination in one element is reversed in the other (S/F).

The fixed active, switched parasitic antenna (FASPA) has two symmetrically located parasitic antennas located either side of the active element marked as F in Fig. 8.16b.

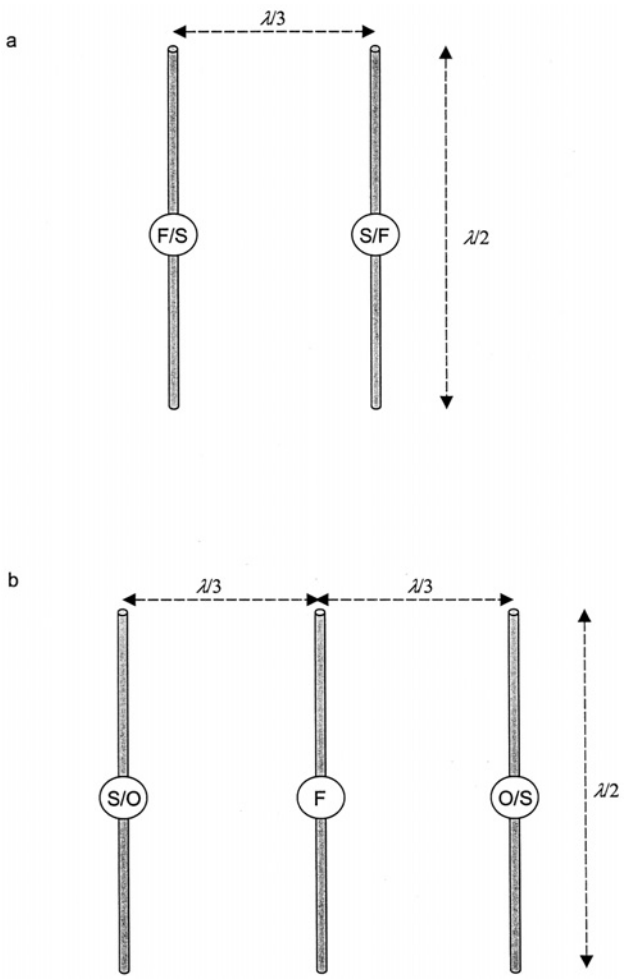


Figure 8.16 Switched parasitic antennas. (a) SASPA using two dipole elements: F/S allows the center of the element to be switched between an RF feed point and a short-circuit (parasitic) element. (b) FASPA using a fixed feed element (F) and two parasitic elements that can be switched between short and open circuit alternatively (S/O).

© 2006 by Taylor & Francis Group, LLC

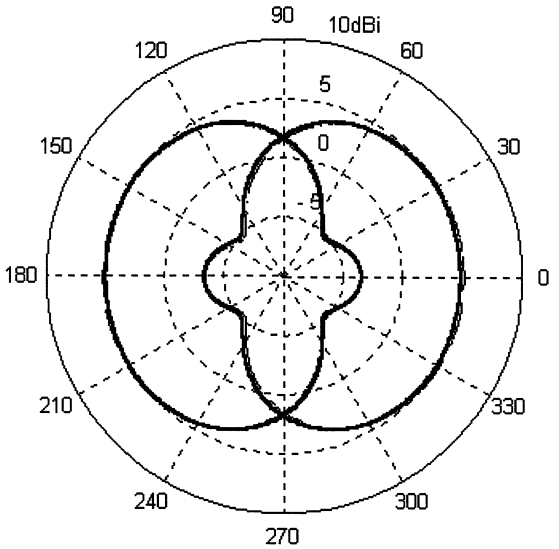


Figure 8.17 Radiation pattern for switched parasitic antennas shown in Fig. 8.16.

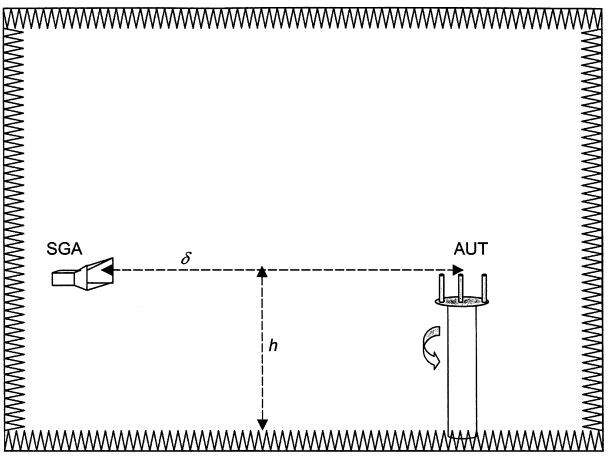


Figure 8.18 Anechoic chamber test facility for measuring antenna radiation patterns. The walls are lined with pyramidal absorber. The antenna under test (AUT) is rotated around a vertical axis while illuminated by the standard gain antenna (SGA).

One parasitic element switch is set to open circuit and the other to short circuit. This is shown as S/O and O/S in Fig. 8.16b. The radiation is directed away from the short-circuit element. When the switch settings are reversed, the principal direction is reversed. Note that in both cases (i.e., FASPA and SASPA), the input impedance of the feed point is independent of the direction of the main beam, and the beam width and null position remains constant relative to the main beam. The switching can be achieved using active devices (e.g., FET or p.i.n. diodes) or passive switches (e.g., RF MEMs switches). While the switches are driven by a DC voltage, the impedance of the switch at open and short circuit must provide sufficient RF isolation. This can limit the frequency of operation of switched parasitic antennas.

© 2006 by Taylor & Francis Group, LLC

It is possible to arrange a continuous rather than discrete switching operation by changing the capacitive load at the center of the parasitic elements. This can be achieved using variable reactive loads [8]. This allows for a more continuous scan of the beam, but with potentially increased variation of the input impedance of the antenna.

Smart antennas must be controlled using a computer or microprocessor in a feedback loop. The SNI ratio into the detector needs to be determined during a scan of all possible main beam directions (a global scan), and then the signal source tracked using a dithering procedure (i.e., by checking the SNI in adjacent main beam positions) together with prediction techniques should the source or receiving antenna be moving. The response time of the smart antenna is dependent on the switching speed, the detector settling speed and the time required to verify the identification of the required source [5].

Two-dimensional phased arrays can have a main beam variation in both θ and ϕ directions, while the switched parasitic antennas are most commonly limited to control in one angular direction. The SNI is sometimes achieved with a large beamwidth antenna by positioning a sharp null in the direction of a strong interfering noise source rather than simply seeking the strongest signal. In a 2D phased array, the search procedure can be quite complex and the computation time required for the calculations may be significant. This is an active research area for digital processing specialists. Direction finding algorithms include the MUSIC technique used for multiple source location and tracking.

8.6. ANTENNA MEASUREMENTS

Following antenna fabrication, the radiation $F(\phi,\theta)$ and impedance $Z(f)$ descriptions of an antenna system usually must be verified experimentally. While there is now an extensive number of computer modeling packages and techniques that allow the accurate calculation of these parameters [3], variations in machining tolerances, the effects of finite conductivity and dielectric loss, and increasing problems associated with the spurious generation of intermodulation frequencies in adjacent frequency bands ensure that antenna testing remains very important.

When planning antenna measurements, there are several important factors to be considered [9]. The antenna performance must be isolated from its supporting structures unless they are to be part of the final installation environment. This commonly means that all objects (both conductive and dielectric) must be located away from the antenna by several Fresnel zones [1]. The first Fresnel zone is defined as the region of space in which the direct radiation path and any possible reflection path differ in length by less than $\lambda/2$. The second zone has reflections from a distance greater than $\lambda/2$ and less than λ. Higher order Fresnel zones follow this definition with the reflection path length increasing by $\lambda/2$ for the next Fresnel zone. Clearly, objects located in the near field of the radiating structure will influence both $F(\phi,\theta)$ and $Z(f)$. This general principle applies to ground reflections, side wall and ceiling reflections, in addition to the effects of feed cables to both the transmitter and receiver, and the antennas themselves. For this reason, for an antenna with a maximum aperture dimension of $D > \lambda$, the separation distance between the two antennas d must be greater than that given by the equation [9]

$$d > \frac{2D^2}{\lambda} \tag{8.9}$$

© 2006 by Taylor & Francis Group, LLC

noting that D must be the maximum aperture dimension for both antennas. Secondly, the antennas should be at the same height h above a reflecting ground plane where [9]

$$h > 4D \tag{8.10}$$

if the antenna is to be rotated about a vertical axis (see Fig. 8.18).

Given these two restrictions, antenna measurements are generally made in an open range (i.e., no obstacles apart from the ground for many Fresnel zones) or in a chamber lined with EM absorbing materials—an anechoic chamber. A number of such materials are available commercially including pyramidal cones made from carbon impregnated foam [2] and flat ceramic tiles. Both are limited to particular frequency bands. In the case of radiation patterns with very large differences between the lobes and nulls (e.g., high-gain antennas), the finite reflections from these absorbing tiles can affect results significantly.

The gain of an antenna can be determined using two standard gain antennas (SGA) and the antenna under test (AUT), or two identical AUT's. The procedure requires the feed cables to both the transmitting antenna and the receiving antenna to be disconnected from the antennas and shorted together. The vector network analyzer is then calibrated for zero insertion loss and the received power P_{sc} noted for every frequency of interest. The two identical antennas (SGAs or AUTs) are separated by distance d_1 where $d_1 > d$ and d is defined by Eq. (8.9), and the received power P is again noted. The free-space path, loss P_L is given by the expression

$$P_L = 20 \log \frac{2\pi d_1}{\lambda} \tag{8.11}$$

If all values are in dB, the gain of the identical antennas G_i in dB is given by the expression

$$G_i = \frac{P + P_L - P_{sc}}{2} \tag{8.12}$$

In the standard gain horn measurement, the gain of the AUT (G_a) is measured from a power measurement taken from the SGA $\Longleftrightarrow$ AUT measurement P_d using the following calculation:

$$G_a = P + P_L - P_{sc} - G_{sc} \tag{8.13}$$

where G_{sc} is the gain of the SGA. These gain determinations include the antenna mismatch S_{11} for both antennas. A simple, yet very sensitive, test to verify that the antennas are identical is to compare the $S_{11}(f)$ for both antennas.

The input impedance of the antenna is calculated with the antenna in free space or in an anechoic chamber. Initially the effects of the cable must be removed from the calculation. This requires a three-point calibration procedure across the frequency range of interest. A high quality short-circuit, open-circuit, and matched load termination must be sequentially applied to the vector network analyzer (VNA) and the reflected field strength (amplitude and phase) noted at every frequency. These measurements can be designated Γ_s, Γ_o, and Γ_0, respectively. The AUT is then attached to the end of the cable, and Γ_a measured. The impedance of the antenna is then determined mathematically.

© 2006 by Taylor & Francis Group, LLC

The accuracy of the measurement depends strongly on the quality of the loads used in the calibration procedure in addition to the noise environment surrounding the antenna [2].

Commonly the calibration procedures are an intrinsic part of the operation of a VNA, and the user is prompted for the appropriate connections during the calibration procedure.

REFERENCES

1. *IEEE Standard Definitions of Terms for Antennas*; IEEE Press: Piscataway, NJ, IEEE Std 145-1993.
2. Balanis, C.A. *Antenna Theory Analysis and Design*; 2nd Ed.; Wiley: New York, 1997.
3. Stutzman, W.L.; Thiele, G.A. *Antenna Theory and Design*; 2nd Ed.; Wiley: New York, 1998.
4. Kraus, J.D. *Antennas*; 2nd Ed.; McGraw-Hill: New York, 1988.
5. Thiel, D.V.; Smith, S.A. *Switched Parasitic Antennas for Cellular Communications*; Artech House: Boston MA, 2001.
6. Brown, E.R. RF-MEMs switches for reconfigurable integrated circuits. IEEE Trans. Microwave Theory and Techniques **1998**, *46*, 1868–1880.
7. Goldsmith, C.L.; Yao, Z.; Eshelman, S.; Denniston, D. Performance of low-loss RF-MEMs capacitive switches. IEEE Microwave Guided Wave Lett. **1998**, *22*, 269–271.
8. Harrington, R.F. Reactively controlled directive arrays. IEEE Trans. Antennas Propag. **1978**, *26*, 390–395.
9. IEEE Antenna Standards Committee. IEEE standard test procedures for antennas. ANSI/IEEE Std 149-1979.

© 2006 by Taylor & Francis Group, LLC

9
Antennas: Representative Types

David R. Jackson, Jeffery T. Williams, and Donald R. Wilton
University of Houston
Houston, Texas

In Chapter 8, an overview of basic antenna terminology and antenna properties was given, including a discussion of concepts that are common to all antennas. In the present chapter, the discussion is focused on the specific properties of several representative classes of antennas that are commonly used. These include microstrip antennas, broadband antennas, phased arrays, traveling and leaky-wave antennas, and aperture antennas. Of course, in a single chapter, it is impossible to cover all of the types of antennas that are commonly used, or even to adequately cover all of the design aspects of any one type of antenna. However, this chapter should provide enough information about the five types of antennas that are discussed here to allow the reader to obtain a basic overview of the fundamental properties of these major classes of antennas and to see how the properties vary from one class to another. The references that are provided can be consulted to obtain more detailed information about any of these types of antennas or to learn about other types of antennas not discussed here.

9.1. MICROSTRIP ANTENNAS

9.1.1. Introduction

Microstrip antennas are one of the most widely used types of antennas in the microwave frequency range, and they are often used in the millimeter-wave frequency range as well [1–3]. (Below approximately 1 GHz, the size of a microstrip antenna is usually too large to be practical, and other types of antennas such as wire antennas dominate.) Also called *patch* antennas, microstrip patch antennas consist of a metallic patch of metal that is on top of a grounded dielectric substrate of thickness h, with relative permittivity and permeability ε_r and μ_r as shown in Fig. 9.1 (usually $\mu_r = 1$). The metallic patch may be of various shapes, with rectangular and circular being the most common, as shown in Fig. 9.1. Most of the discussion in this section will be limited to the rectangular patch, although the basic principles are the same for the circular patch. (Many of the CAD formulas presented will apply approximately for the circular patch if the circular patch is modeled as a square patch of the same area.) Various methods can be used to feed the patch, as discussed below.

© 2006 by Taylor & Francis Group, LLC

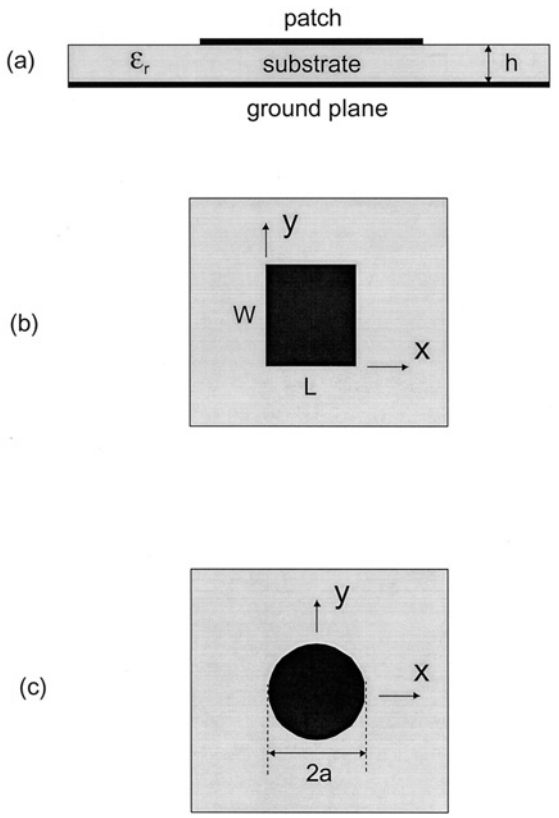


Figure 9.1 Geometry of microstrip patch antenna: (a) side view showing substrate and ground plane, (b) top view showing rectangular patch, and (c) top view showing circular patch.

One advantage of the microstrip antenna is that it is usually low profile, in the sense that the substrate is fairly thin. If the substrate is thin enough, the antenna actually becomes "conformal," meaning that the substrate can be bent to conform to a curved surface (e.g., a cylindrical structure). A typical substrate thickness is about $0.02\lambda_0$. The metallic patch is usually fabricated by a photolithographic etching process or a mechanical milling process, making the construction relatively easy and inexpensive (the cost is mainly that of the substrate material). Other advantages include the fact that the microstrip antenna is usually lightweight (for thin substrates) and durable.

Disadvantages of the microstrip antenna include the fact that it is usually narrowband, with bandwidths of a few percent being typical. Some methods for enhancing bandwidth are discussed later, however. Also, the radiation efficiency of the patch antenna tends to be lower than those of some other types of antennas, with efficiencies between 70% and 90% being typical.

9.1.2. Basic Principles of Operation

The metallic patch essentially creates a resonant cavity, where the patch is the top of the cavity, the ground plane is the bottom of the cavity, and the edges of the patch form the sides of the cavity. The edges of the patch act approximately as an open-circuit boundary

© 2006 by Taylor & Francis Group, LLC

condition. Hence, the patch acts approximately as a cavity with perfect electric conductor on the top and bottom surfaces, and a perfect "magnetic conductor" on the sides. This point of view is very useful in analyzing the patch antenna, as well as in understanding its behavior. Inside the patch cavity the electric field is essentially z directed and independent of the z coordinate. Hence, the patch cavity modes are described by a double index (m, n). For the (m, n) cavity mode of the rectangular patch in Fig. 9.1b, the electric field has the form

$$E_z(x, y) = A_{mn} \cos\frac{m\pi x}{L} \cos\frac{n\pi y}{W} \tag{9.1}$$

where L is the patch length and W is the patch width. The patch is usually operated in the (1, 0) mode, so that L is the resonant dimension, and the field is essentially constant in the y direction. The surface current on the bottom of the metal patch is then x directed, and is given by

$$J_{sx}(x) = A_{10}\left(\frac{\pi/L}{j\omega\mu_0\mu_r}\right) \sin\left(\frac{\pi x}{L}\right) \tag{9.2}$$

For this mode the patch may be regarded as a wide microstrip line of width W, having a resonant length L that is approximately one-half wavelength in the dielectric. The current is maximum at the center of the patch, $x = L/2$, while the electric field is maximum at the two "radiating" edges, $x = 0$ and $x = L$. The width W is usually chosen to be larger than the length ($W = 1.5\,L$ is typical) to maximize the bandwidth, since the bandwidth is proportional to the width. [The width should be kept less than twice the length, however, to avoid excitation of the (0, 1) mode.]

At first glance, it might appear that the microstrip antenna will not be an effective radiator when the substrate is electrically thin, since the patch current in Eq. (9.2) will be effectively shorted by the close proximity to the ground plane. If the modal amplitude A_{10} were constant, the strength of the radiated field would in fact be proportional to h. However, the Q of the cavity therefore increases as h decreases (the radiation Q is inversely proportional to h). Therefore, the amplitude A_{10} of the modal field at resonance is inversely proportional to h. Hence, the strength of the radiated field from a resonant patch is essentially independent of h, if losses are ignored. The resonant input resistance will likewise be nearly independent of h. This explains why a patch antenna can be an effective radiator even for very thin substrates, although the bandwidth will be small.

9.1.3. Feeding Techniques

The microstrip antenna may be fed in various ways. Perhaps the most common is the direct probe feed, shown in Fig. 9.2a for a rectangular patch, where the center conductor of a coaxial feed line penetrates the substrate to make direct contact with the patch. For linear polarization, the patch is usually fed along the centerline, $y = W/2$. The feed point location at $x = x_f$ controls the resonant input resistance. The input resistance is highest when the patch is fed at the edge and smallest (essentially zero) when the patch is fed at the center ($x = L/2$). Another common feeding method, preferred for planar fabrication, is the direct-contact microstrip feed line, shown in Fig. 9.2b. An inset notch is used to control the resonant input resistance at the contact point. The input impedance seen by the

© 2006 by Taylor & Francis Group, LLC

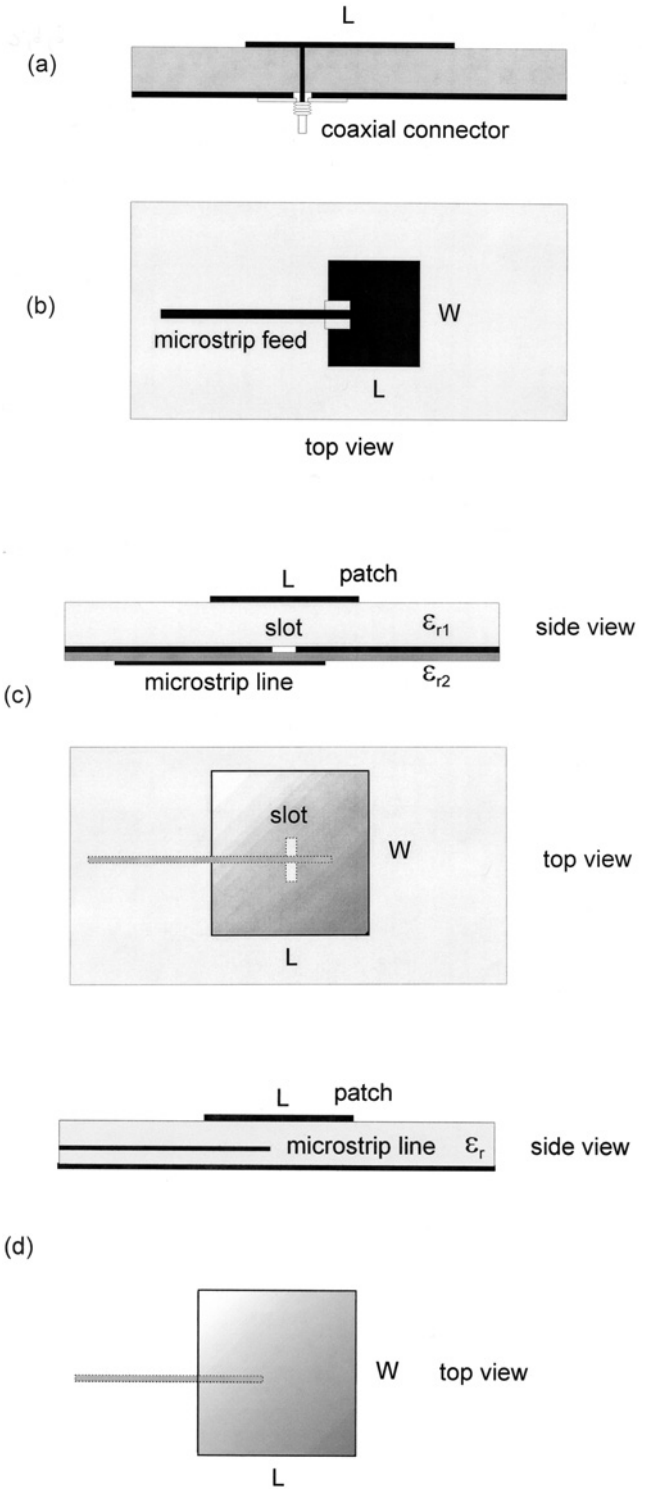


Figure 9.2 Common feeding techniques for a patch antenna: (a) coaxial probe feed, (b) microstrip line feed, (c) aperture-coupled feed, and (d) electromagnetically coupled (proximity) feed.

© 2006 by Taylor & Francis Group, LLC

microstrip line is approximately the same as that seen by a probe at the contact point, provided the notch does not disturb the modal field significantly.

An alternative type of feed is the aperture-coupled feed shown in Fig. 9.2c. In this scheme, a microstrip line on a back substrate excites a slot in the ground plane, which then excites the patch cavity. This scheme has the advantage of isolating the feeding network from the radiating patch element. It also overcomes the limitation on substrate thickness imposed by the feed inductance of a coaxial probe, so that thicker substrates and hence higher bandwidths can be obtained. Using this feeding technique together with a foam substrate, it is possible to achieve bandwidths greater than 25% [4].

Another alternative, which has some of the advantages of the aperture-coupled feed, is the "electromagnetically coupled" or "proximity" feed, shown in Fig. 9.2d. In this arrangement the microstrip line is on the same side of the ground plane as the patch, but does not make direct contact. The microstrip line feeds the patch via electromagnetic (largely capacitive) coupling. With this scheme it is possible to keep the feed line closer to the ground plane compared with the direct feed, in order to minimize feed line radiation. However, the fabrication is more difficult, requiring two substrate layers. Another variation of this technique is to have the microstrip line on the same layer as the patch, with a capacitive gap between the line and the patch edge. This allows for an input match to be achieved without the use of a notch.

9.1.4. Resonance Frequency

The resonance frequency for the $(1, 0)$ mode is given by

$$f_0 = \frac{c}{2L_e\sqrt{\varepsilon_r}} \tag{9.3}$$

where c is the speed of light in vacuum. To account for the fringing of the cavity fields at the edges of the patch, the length, the effective length L_e is chosen as

$$L_e = L + 2\Delta L \tag{9.4}$$

The Hammerstad formula for the fringing extension is [1]

$$\Delta L/h = 0.412\left[\frac{(\varepsilon_{\text{eff}} + 0.3)((W/h) + 0.264)}{(\varepsilon_{\text{eff}} - 0.258)((W/h) + 0.8)}\right] \tag{9.5}$$

where

$$\varepsilon_{\text{eff}} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left(1 + 10\frac{h}{W}\right)^{-1/2} \tag{9.6}$$

9.1.5. Radiation Patterns

The radiation field of the microstrip antenna may be determined using either an "electric current model" or a "magnetic current model." In the electric current model, the current

© 2006 by Taylor & Francis Group, LLC

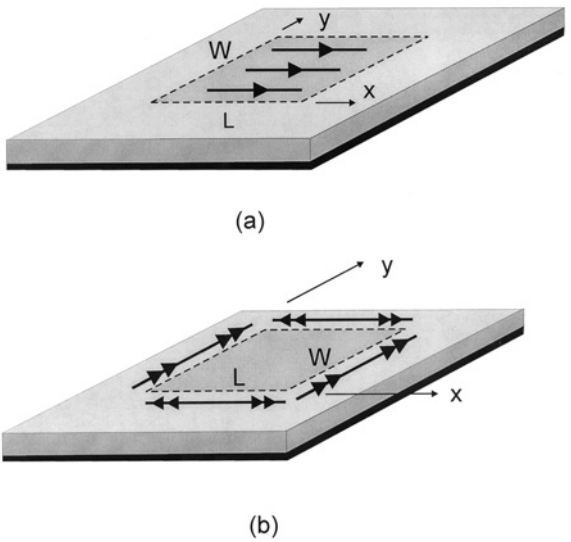


Figure 9.3 Models that are used to calculate the radiation from a microstrip antenna (shown for a rectangular patch): (a) electric current model and (b) magnetic current model.

in Eq. (9.2) is used directly to find the far-field radiation pattern. Figure 9.3a shows the electric current for the $(1, 0)$ patch mode. If the substrate is neglected (replaced by air) for the calculation of the radiation pattern, the pattern may be found directly from image theory. If the substrate is accounted for, and is assumed infinite, the reciprocity method may be used to determine the far-field pattern [5].

In the magnetic current model, the equivalence principle is used to replace the patch by a magnetic surface current that flows on the perimeter of the patch. The magnetic surface current is given by

$$\mathbf{M}_s = -\hat{\mathbf{n}} \times \mathbf{E} \tag{9.7}$$

where $\mathbf{E}$ is the electric field of the cavity mode at the edge of the patch and $\hat{\mathbf{n}}$ is the outward pointing unit-normal vector at the patch boundary. Figure 9.3b shows the magnetic current for the $(1, 0)$ patch mode. The far-field pattern may once again be determined by image theory or reciprocity, depending on whether the substrate is neglected [5]. The dominant part of the radiation field comes from the "radiating edges" at $x = 0$ and $x = L$. The two nonradiating edges do not affect the pattern in the principal planes (the E plane at $\phi = 0$ and the H plane at $\phi = \pi/2$), and have a small effect for other planes.

It can be shown that the electric and magnetic current models yield exactly the same result for the far-field pattern, provided the pattern of each current is calculated in the presence of the substrate at the resonant frequency of the patch cavity mode [5]. If the substrate is neglected, the agreement is only approximate, with the largest difference being near the horizon.

© 2006 by Taylor & Francis Group, LLC

According to the electric current model, accounting for the infinite substrate, the far-field pattern is given by [5]

$$E_i(r,\theta,\phi) = E_i^h(r,\theta,\phi)\left(\frac{\pi WL}{2}\right)\left[\frac{\sin(k_y W/2)}{k_y W/2}\right]\left[\frac{\cos(k_x L/2)}{(\pi/2)^2-(k_x L/2)^2}\right] \tag{9.8}$$

where

$$k_x = k_0 \sin\theta\cos\phi \tag{9.9}$$

$$k_y = k_0 \sin\theta\sin\phi \tag{9.10}$$

and E_i^h is the far-field pattern of an infinitesimal (Hertzian) unit-amplitude x-directed electric dipole at the center of the patch. This pattern is given by [5]

$$E_\theta^h(r,\theta,\phi) = E_0\cos\phi\, G(\theta) \tag{9.11}$$

$$E_\phi^h(r,\theta,\phi) = -E_0\sin\phi F(\theta) \tag{9.12}$$

where

$$E_0 = \left(\frac{-j\omega\mu_0}{4\pi r}\right)e^{-jk_0 r} \tag{9.13}$$

$$F(\theta) = \frac{2\tan(k_0 hN(\theta))}{\tan(k_0 hN(\theta)) - j(N(\theta)/\mu_r)\sec\theta} \tag{9.14}$$

$$G(\theta) = \frac{2\tan(k_0 hN(\theta))cos\theta}{\tan(k_0 hN(\theta)) - j[\varepsilon_r/N(\theta)]\cos\theta} \tag{9.15}$$

and

$$N(\theta) = \sqrt{n_1^2 - \sin^2(\theta)} \tag{9.16}$$

$$n_1 = \sqrt{\varepsilon_r\mu_r} \tag{9.17}$$

The radiation patterns (E and H plane) for a rectangular patch antenna on an infinite nonmagnetic substrate of permittivity $\varepsilon_r = 2.2$ and thickness $h/\lambda_0 = 0.02$ are shown in Fig. 9.4. The patch is resonant with $W/L = 1.5$. Note that the E-plane pattern is broader than the H-plane pattern. The directivity is approximately 6 dB.

9.1.6. Radiation Efficiency

The radiation efficiency of the patch antenna is affected not only by conductor and dielectric losses, but also by surface-wave excitation—since the dominant TM_0 mode of the grounded substrate will be excited by the patch. As the substrate thickness decreases, the effect of the conductor and dielectric losses becomes more severe, limiting the efficiency. On the other hand, as the substrate thickness increases, the surface-wave power

© 2006 by Taylor & Francis Group, LLC

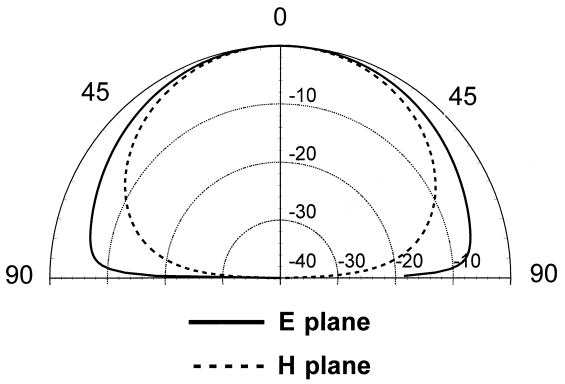


Figure 9.4 The radiation patterns for a rectangular patch antenna on an infinite substrate of permittivity $\varepsilon_r = 2.2$ and thickness $h/\lambda_0 = 0.02$. The patch is resonant with $W/L = 1.5$. The E-plane (xz plane) and H-plane (yz plane) patterns are shown.

increases, thus limiting the efficiency. Surface-wave excitation is undesirable for other reasons as well, since surface waves contribute to mutual coupling between elements in an array and also cause undesirable edge diffraction at the edges of the ground plane or substrate, which often contributes to distortions in the pattern and to back radiation. For an air (or foam) substrate there is no surface-wave excitation. In this case, higher efficiency is obtained by making the substrate thicker, to minimize conductor and dielectric losses (making the substrate too thick may lead to difficulty in matching, however, as discussed above). For a substrate with a moderate relative permittivity such as $\varepsilon_r = 2.2$, the efficiency will be maximum when the substrate thickness is approximately $0.02\lambda_0$.

The radiation efficiency is defined as

$$e_r = \frac{P_{\text{sp}}}{P_{\text{total}}} = \frac{P_{\text{sp}}}{P_{\text{sp}} + P_{\text{c}} + P_{\text{d}} + P_{\text{sw}}} \tag{9.18}$$

where P_{sp} is the power radiated into space, and the total input power P_{total} is given as the sum of P_{c} is the power dissipated by conductor loss, P_{d} is the power dissipated by dielectric loss, and P_{sw} is the surface-wave power. The efficiency may also be expressed in terms of the corresponding Q factors as

$$e_r = \frac{Q_{\text{total}}}{Q_{\text{sp}}} \tag{9.19}$$

where

$$\frac{1}{Q_{\text{total}}} = \frac{1}{Q_{\text{sp}}} + \frac{1}{Q_{\text{sw}}} + \frac{1}{Q_{\text{d}}} + \frac{1}{Q_{\text{c}}} \tag{9.20}$$

© 2006 by Taylor & Francis Group, LLC

The dielectric and conductor Q factors are given by

$$Q_{\mathrm{d}} = \frac{1}{\tan\delta} \tag{9.21}$$

$$Q_{\mathrm{c}} = \frac{1}{2}\eta_0\mu_r\frac{k_0 h}{R_s} \tag{9.22}$$

where $\tan\delta$ is the loss tangent of the substrate and R_s is the surface resistance of the patch and ground plane metal (assumed equal) at radian frequency $\omega = 2\pi f$, given by

$$R_s = \sqrt{\frac{\omega\mu_0}{2\sigma}} \tag{9.23}$$

where σ is the conductivity of the metal.

The space-wave Q factor is given approximately as [6]

$$Q_{\mathrm{sp}} = \frac{3}{16}\left(\frac{\varepsilon_r}{pc_1}\right)\left(\frac{L}{W}\right)\left(\frac{1}{h/\lambda_0}\right) \tag{9.24}$$

where

$$c_1 = 1 - \frac{1}{n_1^2} + \frac{2/5}{n_1^4} \tag{9.25}$$

and

$$p = 1 + \frac{a_2}{10}(k_0 W)^2 + \left(a_2^2 + 2a_4\right)\left(\frac{3}{560}\right)(k_0 W)^4 + c_2\left(\frac{1}{5}\right)(k_0 L)^2 + a_2 c_2\left(\frac{1}{70}\right)(k_0 W)^2(k_0 L)^2 \tag{9.26}$$

with $a_2 = -0.16605$, $a_4 = 0.00761$, and $c_2 = -0.0914153$.

The surface-wave Q factor is related to the space-wave Q factor as

$$Q_{\mathrm{sw}} = Q_{\mathrm{sp}}\left(\frac{e_r^{\mathrm{sw}}}{1 - e_r^{\mathrm{sw}}}\right) \tag{9.27}$$

where e_r^{sw} is the radiation efficiency accounting only for surface-wave loss. This efficiency may be accurately approximated by using the radiation efficiency of an infinitesimal dipole on the substrate layer [6], giving

$$e_r^{\mathrm{sw}} = \frac{1}{1 + (k_0 h)(3/4)(\pi\mu_r)(1/c_1)\left(1 - 1/n_1^2\right)^3} \tag{9.28}$$

A plot of radiation efficiency for a resonant rectangular patch antenna with $W/L = 1.5$ on a nonmagnetic substrate of relative permittivity $\varepsilon_r = 2.2$ or $\varepsilon_r = 10.8$ is shown in Fig. 9.5. The conductivity of the copper patch and ground plane is assumed to be $\sigma = 3.0 \times 10^7$ [S/m]

© 2006 by Taylor & Francis Group, LLC

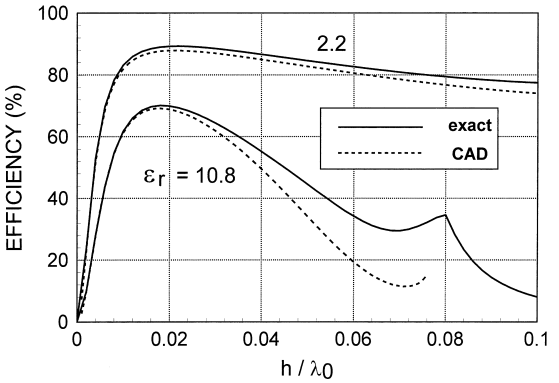


Figure 9.5 Radiation efficiency (%) for a rectangular patch antenna versus normalized substrate thickness. The patch is resonant at 5.0 GHz with $W/L = 1.5$ on a substrate of relative permittivity $\varepsilon_r = 2.2$ or $\varepsilon_r = 10.8$. The conductivity of the copper patch and ground plane is $\sigma = 3.0 \times 10^7$ S/m and the dielectric loss tangent is $\tan\delta = 0.001$. The exact efficiency is compared with the result of the CAD formula [Eq. (9.19) with Eqs. (9.20)–(9.28))].

and the dielectric loss tangent is taken as $\tan\delta = 0.001$. The resonance frequency is 5.0 GHz. [The result is plotted versus normalized (electrical) thickness of the substrate, which does not involve frequency. However, a specified frequency is necessary to determine conductor loss.] For $h/\lambda_0 < 0.02$, the conductor and dielectric losses dominate, while for $h/\lambda_0 < 0.02$, the surface-wave losses dominate. (If there were no conductor or dielectric losses, the efficiency would approach 100% as the substrate thickness approaches zero.)

9.1.7. Bandwidth

The bandwidth increases as the substrate thickness increases (the bandwidth is directly proportional to h if conductor, dielectric, and surface-wave losses are ignored). However, increasing the substrate thickness lowers the Q of the cavity, which increases spurious radiation from the feed, as well as from higher-order modes in the patch cavity. Also, the patch typically becomes difficult to match as the substrate thickness increases beyond a certain point (typically about $0.05\lambda_0$). This is especially true when feeding with a coaxial probe, since a thicker substrate results in a larger probe inductance appearing in series with the patch impedance. However, in recent years considerable effort has been spent to improve the bandwidth of the microstrip antenna, in part by using alternative feeding schemes. The aperture-coupled feed of Fig. 9.2c is one scheme that overcomes the problem of probe inductance, at the cost of increased complexity [7].

Lowering the substrate permittivity also increases the bandwidth of the patch antenna. However, this has the disadvantage of making the patch larger. Also, because the Q of the patch cavity is lowered, there will usually be increased radiation from higher-order modes, degrading the polarization purity of the radiation.

By using a combination of aperture-coupled feeding and a low-permittivity foam substrate, bandwidths exceeding 25% have been obtained. The use of stacked patches (a parasitic patch located above the primary driven patch) can also be used to

© 2006 by Taylor & Francis Group, LLC

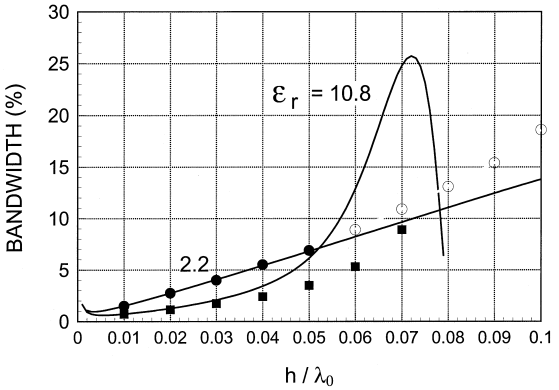


Figure 9.6 Bandwidth (%) for a rectangular patch antenna versus normalized substrate thickness. The parameters are the same as in Fig. 9.5. The exact bandwidth (SWR < 2.0) from a cavity model calculation is compared with the result of the CAD formula [Eq. (9.29)]. The exact calculation assumes a feed location $x_0 = L/4$, $y_0 = W/2$. The exact result is shown with a solid line, and the CAD results are shown with the discrete data points. For the low-permittivity substrate, the hollow dots indicate that the reactance does not go to zero at any frequency. For these cases the resonance frequency is defined as the frequency that minimizes the reactance, and the corresponding minimum reactance value is subtracted from the impedance at each frequency in order to define the SWR bandwidth.

increase bandwidth even further, by increasing the effective height of the structure and by creating a double-tuned resonance effect [8].

A CAD formula for the bandwidth (defined by SWR < 2.0) is

$$BW = \frac{1}{\sqrt{2}}\left(\tan\delta + \frac{R_s}{\pi\eta_0\mu_r}\frac{1}{h/\lambda_0} + \frac{16}{3}\frac{pc_1}{\varepsilon_r}\frac{h}{\lambda_0}\frac{W}{L}\frac{1}{e_r^{sw}}\right) \tag{9.29}$$

where the terms have been defined in the previous section on radiation efficiency. The result should be multiplied by 100 to get percent bandwidth. Note that neglecting conductor and dielectric loss yields a bandwidth that is proportional to the substrate thickness h.

Figure 9.6 shows calculated and measured bandwidth for the same patch in Fig. 9.5. It is seen that bandwidth is improved by using a lower substrate permittivity and by making the substrate thicker.

9.1.8. Input Impedance

Several approximate models have been proposed for the calculation of input impedance for a probe-fed patch. These include the transmission line method [9], the cavity model [10], and the spectral-domain method [11]. These models usually work well for thin substrates, typically giving reliable results for $h/\lambda_0 < 0.02$. Commercial simulation tools using FDTD, FEM, or MoM can be used to accurately predict the input impedance for any substrate thickness. The cavity model has the advantage of allowing for a simple physical CAD model of the patch to be developed, as shown in Fig. 9.7. In this model the patch cavity is modeled as a parallel RLC circuit, while the probe inductance is

© 2006 by Taylor & Francis Group, LLC

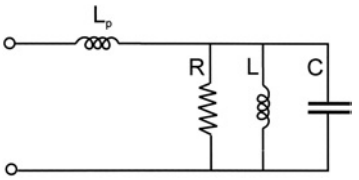


Figure 9.7 CAD model for the input impedance of a coaxial probe-fed microstrip antenna, operating near the resonance frequency.

modeled as a series inductor. The input impedance of this circuit is approximately described by

$$Z_{\text{in}} \approx jX_f + \frac{R}{1 + j2Q(f/f_0 - 1)} \tag{9.30}$$

where f_0 is the resonance frequency, R is the input resistance at the resonance of the RLC circuit (where the input resistance of the patch is maximum), $Q = Q_{\text{total}}$ is the quality factor of the patch cavity (9.20), and $X_f = \omega L_p$ is the feed (probe) reactance of the coaxial probe. A CAD formula for the input resistance R is

$$R = R_{\text{edge}} \cos^2\left(\frac{\pi x_0}{L}\right) \tag{9.31}$$

where the input resistance at the edge is

$$R_{\text{edge}} = \frac{(4/\pi)(\mu_r \eta_0)(L/W)(h/\lambda_0)}{\tan\delta + (R_s/(\pi\eta_0\mu_r))(1/(h/\lambda_0)) + (16/3)(pc_1/\varepsilon_r)(W/L)(h/\lambda_0)\left(1/e_r^{hed}\right)} \tag{9.32}$$

A CAD formula for the feed reactance due to the probe is

$$X_f = \frac{\eta_0}{2\pi}\mu_r(k_0 h)\left[-\gamma + \ln\frac{2}{\sqrt{\varepsilon_r\mu_r}(k_0 a)}\right] \tag{9.33}$$

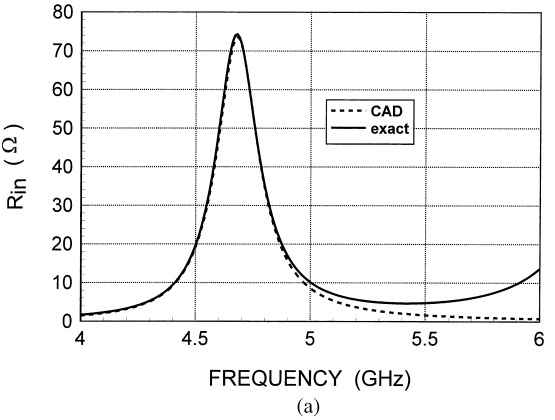
where $\gamma = 0.577216$ is Euler's constant.

Figure 9.8 shows a comparison of the input impedance obtained from the simple CAD model Eq. (9.30) with that obtained by a more accurate cavity model analysis. At the resonance frequency, the substrate thickness is approximately $0.024\lambda_0$. Near the resonance frequency, the simple CAD model gives results that agree quite well with the cavity model.

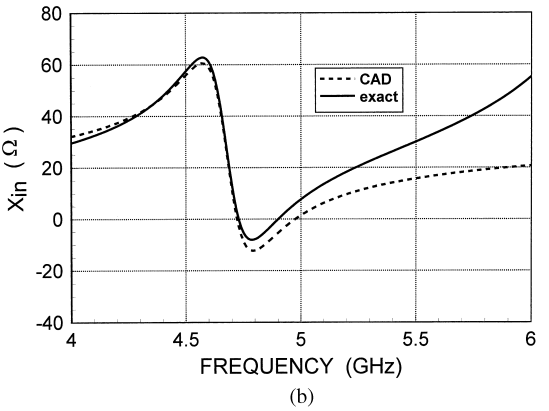
9.1.9. Improving Performance

Much research has been devoted to improving the performance characteristics of the microstrip antenna. To improve bandwidth, the use of thick low-permittivity (e.g., foam) substrates can give significant improvement. To overcome the probe inductance associated with thicker substrates, the use of capacitive-coupled feeds such as the top-loaded probe [12] or the L-shaped probe [13] shown in Fig. 9.9a and b may be used. Alternatively, the aperture-coupled fed shown in Fig. 9.2c may be used, which also has the advantage of

© 2006 by Taylor & Francis Group, LLC



(a)



(b)

Figure 9.8 Input impedance versus frequency for a rectangular coaxial probe-fed patch antenna. The results from the CAD model in Fig. 9.7 are compared with those obtained by a cavity-model analysis: (a) input resistance and (b) input reactance. $L = 2.0$ cm and $W/L = 1.5$. The feed probe is located at $x_0 = L/4$, $y_0 = W/2$ and has a radius of 0.05 cm. The substrate has a permittivity of $\varepsilon_r = 2.2$ and a thickness of 0.1524 cm.

eliminating spurious probe radiation. To increase the bandwidth even further, a stacked patch arrangement may be used, in which a parasitic patch is stacked above the driven patch [8]. This may be done using either a probe feed or, to obtain even higher bandwidths, an aperture-coupled feed (Fig. 9.9c). The bandwidth enhancement is largely due to the existence of a double resonance and, to some extent, to the fact that one of the radiators is further from the ground plane. Bandwidths as large as one octave (2:1 frequency band) have been obtained with such an arrangement. By using a diplexer feed to split the feeding signal into two separate branches, and feeding two aperture-coupled stacked patches with different center frequencies, bandwidths of 4:1 have been obtained [14]. Parasitic patches may also be placed on the same substrate as the driven patch, surrounding the driven patch. A pair of parasitic patches may be coupled to the radiating edges, the nonradiating edges, or all four edges [15]. This planar arrangement saves vertical height and allows for easier fabrication, although the substrate area occupied by the antenna to be larger, and there may be more variation of the radiation pattern across

© 2006 by Taylor & Francis Group, LLC

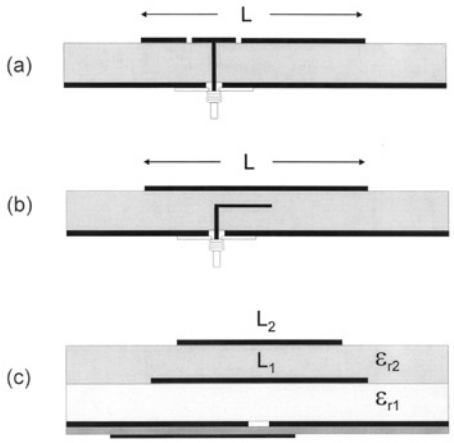


Figure 9.9 Some schemes for improving bandwidth: (a) probe with a capacitive top loading, (b) L-shaped probe, (c) stacked patches, and (d) aperture-coupled stacked patches.

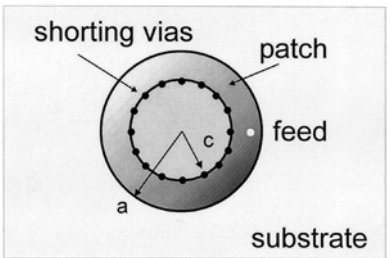


Figure 9.10 The ''reduced-surface-wave'' microstrip antenna. This antenna excites less surface-wave and lateral radiation than does a conventional microstrip antenna. The antenna consists of a circular patch of radius a that has a short-circuit boundary (array of vias) at an inner radius c.

the frequency band since the current distribution on the different patches changes with frequency. Broadbanding may also be achieved through the use of slots cut into the patch, as in the U-slot patch design [16]. This has the advantage of not requiring multiple layers or increasing the size of the patch as with parasitic elements.

Another variation of the microstrip antenna that has been introduced recently is the ''reduced surface wave'' microstrip antenna shown in Fig. 9.10 [17]. This design is a variation of a circular patch, with an inner ring of vias that create a short-circuit inner boundary. By properly selecting the outer radius, the patch excites very little surface-wave field and also only a small amount of lateral (horizontally propagating) radiation. The inner short-circuit boundary is used to adjust the dimensions of the patch cavity (between the inner and outer boundaries) to make the patch resonant. The reduced surface-wave and lateral radiation result in less edge diffraction from the edges of the supporting ground plane, giving smoother patterns in the front-side region and less radiation in the back-side region. Also, there is less mutual coupling between pairs of such antennas, especially as the separation increases. The disadvantage of this antenna is that it is physically fairly large, being about $0.60\lambda_0$ in diameter, regardless of the substrate permittivity.

© 2006 by Taylor & Francis Group, LLC

9.2. BROADBAND ANTENNAS

9.2.1. Introduction

Until relatively recently, broadband antennas (for the purpose of this discussion, broadband suggests bandwidths of approximately an octave, 2:1, or more) have been predominately employed in radar and tracking applications and in specialized broadband communications systems. However, with the move to digital modulation and spread spectrum coding schemes over multiple frequency bands in modern communication systems, the need for broadband antennas has increased rapidly. There are many ways to achieve wideband antenna performance. Typically, however, antennas that provide broadband coverage fall into one of two categories: multiband elements and arrays that simultaneously cover multiple "spot" (narrow) bands, and naturally broadband (quasi-frequency independent) radiators. The focus of this discussion will be on the latter. In addition, the antenna designs considered will be primarily for RF and microwave applications; however, many of the designs can be used at lower and higher frequencies. The discussion will be limited to outlining the general properties and operation of the most common broadband antenna elements; helical, spiral, and log-periodic antennas.

9.2.2. Helical Antenna

Helical antennas, or helixes, are relatively simple structures with one, two, or more wires each wound to form a helix, usually backed by a ground plane or shaped reflector and driven with an appropriate feed [18–20]. The most common design is a single wire (monofilar helix), backed by a ground and fed with a coaxial line, as shown in Fig. 9.11. For this typical helix geometry, L is the axial length, D is the diameter, S is the inner winding spacing, C is the circumference, α is the pitch angle (defined as the angle between a tangent line to the helix wire and the plane perpendicular to the axis of the helix), and a is the radius of the helix wire. The helix has N turns. In general, the radiation properties of the helical antenna are associated with the electrical size of the structure, where the input impedance is more sensitive to the pitch and wire size. Helical antennas have two predominant radiation modes, the normal (broadside) mode and the axial (end-fire) mode. The normal mode occurs when C is small compared to a wavelength and the axial mode occurs when C is on the order of a wavelength. For most applications, the axial mode is used. Hence, the following discussion will focus on the end-fire mode of operation for a helical antenna.

The radiation pattern for the axial (end-fire) mode is characterized by a major lobe along the axial direction. The polarization along this direction is elliptical, and when appropriately designed ($3/4\lambda_0 < C < 4/3\lambda_0$, $S \approx \lambda_0/4$), good circular polarization (CP) can be obtained. The handedness of the radiation is determined by twist of the helix. If wound as a right-handed (RH) screw, the polarization of the radiated field is RH. If wound as a left-handed (LH) screw, the polarization of the radiated field is LH. The helix shown in Fig. 9.11 is LH. The helix is characterized by an approximately real input impedance over a slightly less than 2:1 bandwidth. The value of this impedance ranges between 100 and 200 Ω. For axial mode helix antennas with $C/\lambda_0 \approx 1$ [18],

$$R_{IN} \approx 140 \frac{C}{\lambda_0} \quad [\Omega] \tag{9.34}$$

© 2006 by Taylor & Francis Group, LLC

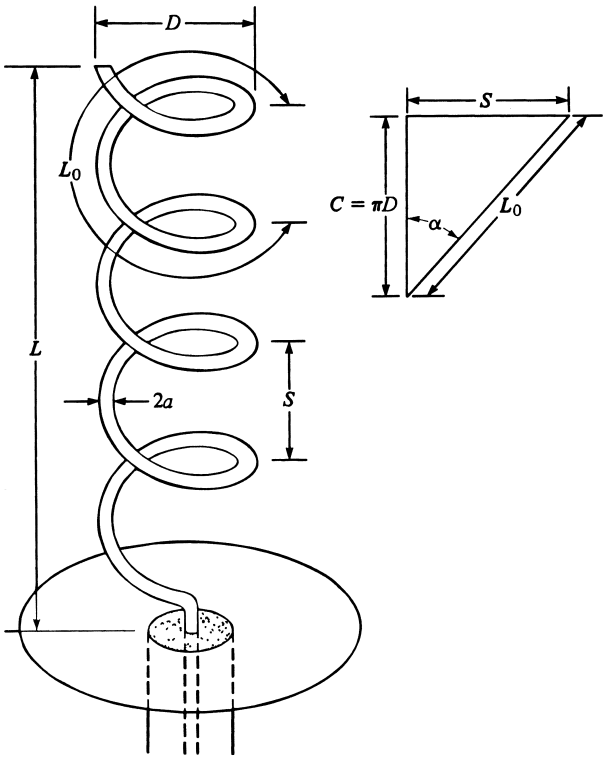


Figure 9.11 Monofilar helical antenna with ground plane [38].

Since most of the feeding coaxial lines have a characteristic impedance significantly less than R_{IN} (typically, 50 Ω), helical antennas are not usually fed directly by the coax as depicted in Fig. 9.11. A variety of techniques have been developed to match the higher impedance helix with the feed coax, including varying the pitch and the diameter of the helix wire at the feed to essentially form a tapered matching section [20]. However, the most common matching technique is to move the coax feed off the axis of the helix and insert a microstrip matching transformer between the coax feed and the beginning of the helix.

The radiation pattern for an axial-mode helix is approximated by treating the helix as a linear-end-fire array of one-wavelength circumference loop antennas with spacing S [21]. The progressive phase shift between the elements corresponds to the shift associated with a traveling wave along the helical wire. The interelement phase shift (δ) is given by

$$\delta \approx -\frac{\omega \ell}{v_p} = -\frac{k_0 L_0}{v_p/c} \tag{9.35}$$

where k_0 is the free-space wavenumber, L_0 is the length along one turn of the helix, c is the speed of light in free space, and v_p phase velocity of the traveling wave current along the helix. For an axial mode helix,

$$\frac{v_p}{c} \approx \frac{L_0/\lambda_0}{S/\lambda_0 + (2N+1)/2N} \tag{9.36}$$

© 2006 by Taylor & Francis Group, LLC

The traveling wave current is a slow wave with respect to c. The current along the helical wire decays away from the feed due to radiation; however, a simplifying assumption used to approximate the radiation pattern is to assume the amplitude of the current on each loop is the same. Hence, the circularly polarized electric field radiated by an axial-mode helix is approximated as

$$E(\theta) \approx A \cos\theta \frac{\sin(N\psi/2)}{N \sin(\psi/2)} \tag{9.37}$$

where $\psi = k_0 S \cos\theta + \delta$. An example of the pattern for an axial-mode helical antenna is shown in Fig. 9.12. The directivity of a helix is approximated by [18]

$$D \approx 15\left(\frac{C}{\lambda_0}\right)^2 \frac{L}{\lambda_0} = 15\left(\frac{C}{\lambda_0}\right)^2 \frac{NS}{\lambda_0} \tag{9.38}$$

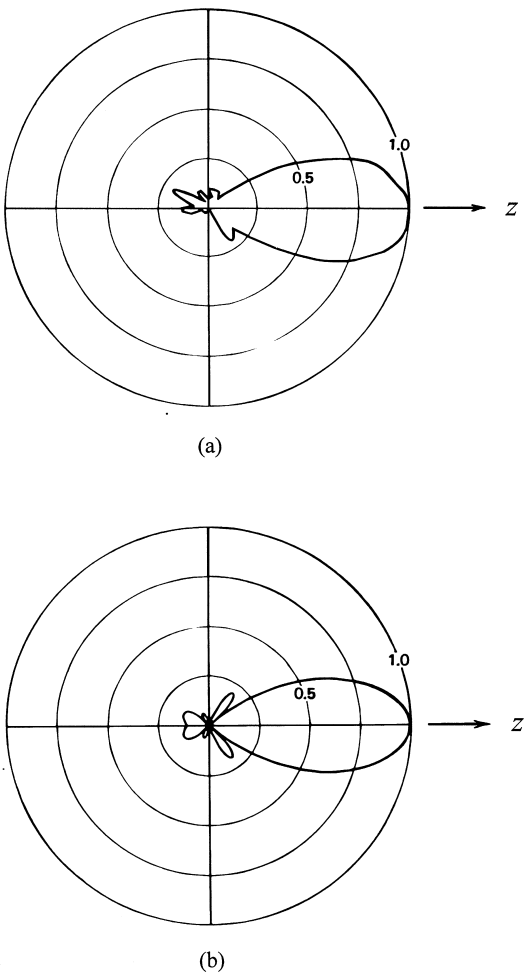


Figure 9.12 Radiation pattern for an axial-mode helical antenna ($N = 10$, $C = \lambda_0, \alpha = 13^\circ$, $f = 8\,\text{GHz}$): (a) measured and (b) theory [25].

© 2006 by Taylor & Francis Group, LLC

To achieve the desired radiation and polarization characteristics the length of the helical antenna needs to be sufficiently large to ensure that the outward propagating traveling wave on the helix is attenuated to the point that the reflected wave at the end of the helix is small. Typically, this is achieved by making $L > \lambda_0/2$. Gradually reducing the radius of the helix turns near the end of the antenna also has the effect of reducing the reflection of the outward traveling wave and helps flatten the impedance characteristics of the antenna [22]. Ideally, the polarization of a helical antenna is circular; however, this is only true for an infinite helix. The polarization of a finite helix is actually elliptical, with an axial ratio approximated by [18]

$$\mathrm{AR} = \frac{2N+1}{2N} \tag{9.39}$$

Thus, as the number of turns is increased (the length of antenna increased for a fixed pitch), the polarization approaches circular.

Multifilar helical antennas typically are slightly more directive and have better axial ratios compared to equivalent monofilar helixes. The feeds for these structures, however, are more complicated [19].

9.2.3. Frequency Independent Antennas

For bandwidths much larger than an octave, (quasi-) frequency independent antennas are typically used. The design of frequency independent antennas is based upon the knowledge that the impedance and radiation properties of an antenna are associated with the electrical dimensions of the structure (dimensions expressed in wavelengths). Hence, if an arbitrary scaling of the antenna structure results in the same structure, possibly rotated about the vertex, the electrical properties of the antenna will be independent of frequency [23]. Such antennas can be described solely by angular specifications. Antennas that can be described on conical surfaces and by equiangular spiral curves satisfy this angle requirement. Theoretically, the structure must be infinite in extent and emanate from a point vertex to be truly frequency independent. In practice, the operating bandwidth of these antennas are limited at low frequencies by the outer dimensions of the structure and how the structure is terminated along the outside boundary. The currents on these antennas tend to decay rapidly away from the center of the structure, particularly beyond the active or radiation region (the portion of the antenna near the radius $r = \lambda/2\pi$) [24]. Thus, if the structure is appropriately truncated beyond the radiation region where the currents are relatively low, the performance of the antenna is not adversely affected. The high-frequency limit of these antennas is dictated by the inside dimensions and the precision of the antenna near the vertex. This is commonly the feed region of these antennas.

It is also interesting to note that the input impedance of a self-complementary antenna is frequency independent. A self-complementary planar structure (as defined by Babinet's principle) is one that remains the same, with the exception of a rotation about the vertex, when the metallic and nonmetallic regions on the antenna surface are interchanged. In this case the input impedance for both structures is the same. If the structures are infinite, the input impedance is independent of frequency and, from Babinet's principle, equal to $188.5\,\Omega$ [25]. The input impedance for a truncated self-complementary antenna is typically slightly less than this value but in practice can be made relatively constant over a wide band. While self-complementary antennas do have a

© 2006 by Taylor & Francis Group, LLC

frequency independent input impedance, being self-complementary is not a necessary requirement for frequency independent performance. Many frequency independent antenna designs are not self-complementary.

9.2.4. Spiral Antennas

Broadband spiral antennas can be realized on either planar or conical surfaces. Planar spirals are typically bidirectional, radiating near circularly polarized fields on both the front and back sides. To eliminate back-side radiation planar spirals are backed by an absorber-filled cavity (more typical) or a conducting ground plane (less typical). Conical spirals are unidirectional, radiating along the direction of the apex, thereby eliminating the requirement for a backing cavity. The most common spiral designs have two arms; however, to improve pattern symmetry and for direction-finding and tracking systems, multiarm (typically four-arm) spirals are used [26].

The planar equiangular spiral antenna (log-spiral) is defined by the equiangular spiral curve shown in Fig. 9.13 [27]

$$r = \rho = \rho_i e^{a(\phi - \phi_i)} \qquad \text{for } \theta = \frac{\pi}{2} \tag{9.40}$$

where ρ is the radial distance from the vertex in the $\theta = \pi/2$ plane, ϕ is the angle from the x axis, ρ_i and ϕ_i are the coordinates to the start of the spiral curve, and a is the flare rate of the curve. As shown in the figure, the flare rate is related to the pitch angle of the spiral (ψ) by

$$\tan \psi = \frac{1}{a} \tag{9.41}$$

Note that beyond the starting point this curve is defined solely by the angle ϕ.

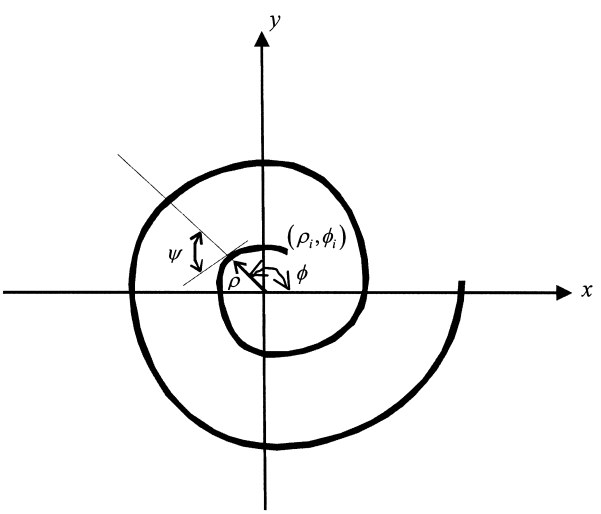


Figure 9.13 Equiangular spiral curve.

© 2006 by Taylor & Francis Group, LLC

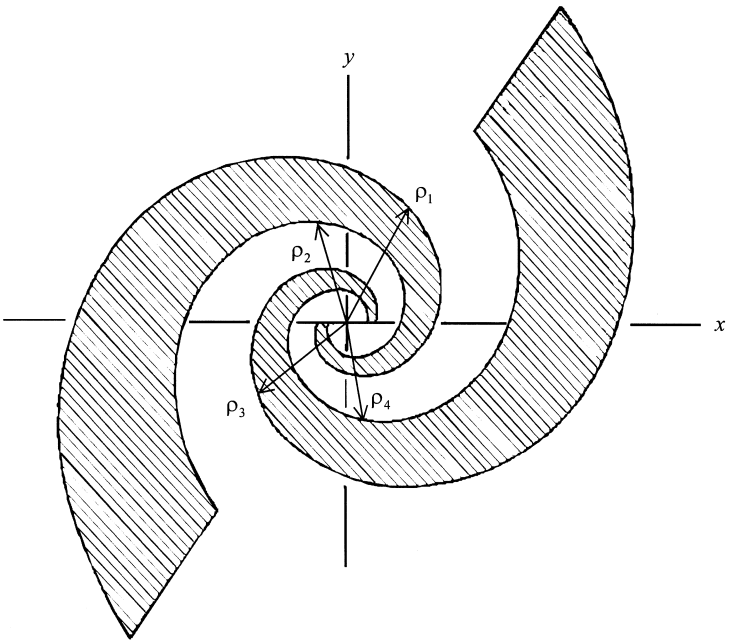


Figure 9.14 Planar two-arm equiangular spiral antenna.

A log-spiral antenna defined using Eq. (9.40) is shown in Fig. 9.14. The first arm, which begins along the x axis in this example, is defined by the edges

$$\begin{aligned} &\text{Outer edge:} \quad \rho_1 = \rho_0 e^{a\phi} \\ &\text{Inner edge:} \quad \rho_2 = \rho_0 e^{a(\phi-\delta)} \end{aligned} \tag{9.42}$$

where

$$\frac{\rho_2}{\rho_1} = e^{-a\delta} < 1 \tag{9.43}$$

The second arm is realized by simply rotating the first arm by π radians. The expressions for the edges of this arm are obtained by multiplying the expressions in Eq. (9.42) by $e^{-a\pi}$. This two-arm structure is self-complementary when $\delta = \pi/2$. The scaling ratio for an equiangular spiral,

$$\tau = \frac{\rho(\phi)}{\rho(\phi + 2\pi)} = e^{-2\pi a} \tag{9.44}$$

should typically be between 0.1 and 0.9. Optimum performance for the equiangular spiral is obtained when the number of turns is between 1.5 and 3. A complementary structure, a spiral slot antenna has similar electrical properties and in many cases is easier to physically implement.

© 2006 by Taylor & Francis Group, LLC

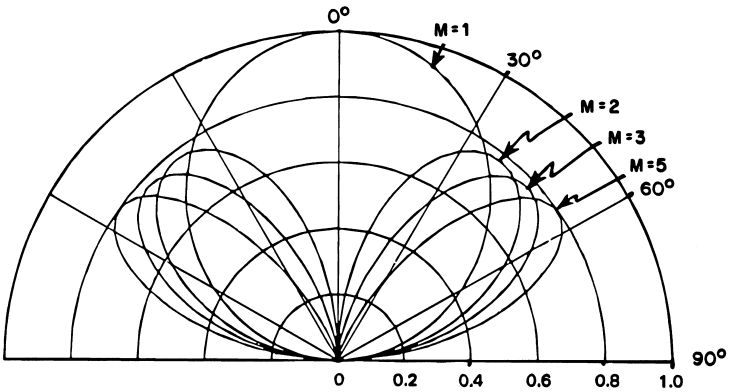


Figure 9.15 Representative radiation pattern for the $M = 1$, 2, 3, and 5 modes of a planar two-arm equiangular spiral antenna.

An N-arm spiral antenna with discrete rotational symmetry characterized by the angle $2\pi/N$ radians supports N normal modes. For the Mth mode ($M \in [1, N]$), the spiral arms are excited with equal amplitudes and a progressive phase shift equal to $e^{-j2\pi M/N}$. Qualitatively, the predominant radiation for the Mth mode is from the active (radiation) region, the region on the spiral near the radius $r = M\lambda/2\pi$. In the active region, the phasing of the currents along the adjacent arms are such that they essentially form an annular ring of traveling-wave current M wavelengths in circumference [24]. All the modes are bidirectional, and all but the $M = 1$ mode have nulls broadside to the antenna. The $M = 1$ mode has single lobes broadside to the front and back sides and is approximately characterized by $\cos\theta$. An example of these patterns is shown in Fig. 9.15. If fed appropriately, the radiation is nearly perfectly circularly polarized (CP) along the axis of the antenna, degrading as the observation angle moves away from broadside. In addition, since physically scaling the equiangular structure is equivalent to a rotation in ϕ, the radiation patterns rotate with frequency. Only at frequencies scaled by τ are the scaled structures congruent and the patterns identical, assuming the arms are appropriately fed and terminated.

Spiral antennas are inherently balanced structures, thereby requiring a balanced feed. Since the antenna is typically fed with a 50 Ω coax and the input impedance for the spiral is on the order of 150–180 Ω, broadband impedance transformers and baluns are required. While beyond the scope of this discussion, the balun design is critical to exciting purely the desired mode on the spiral at all frequencies of operation [26]. If additional modes are excited, the result is typically a degradation in the impedance and CP performance, along with squint in the radiation pattern. In addition, if the currents are not sufficiently attenuated through the active region of the antenna, the residual energy they carry must be either radiated or dissipated to prevent reflections from the end of the arms and the consequent opposite-handed reradiation. Tapering of the ends of the arms and use of resistive–absorbing loads along the outer portion of the antenna are techniques used to damp these residual currents.

Although not strictly frequency independent because the structure is not defined solely by angles, the archimedean spiral design is widely used. Archimedean spirals have bandwidths over 10 : 1 and excellent polarization and pattern characteristics [26].

© 2006 by Taylor & Francis Group, LLC

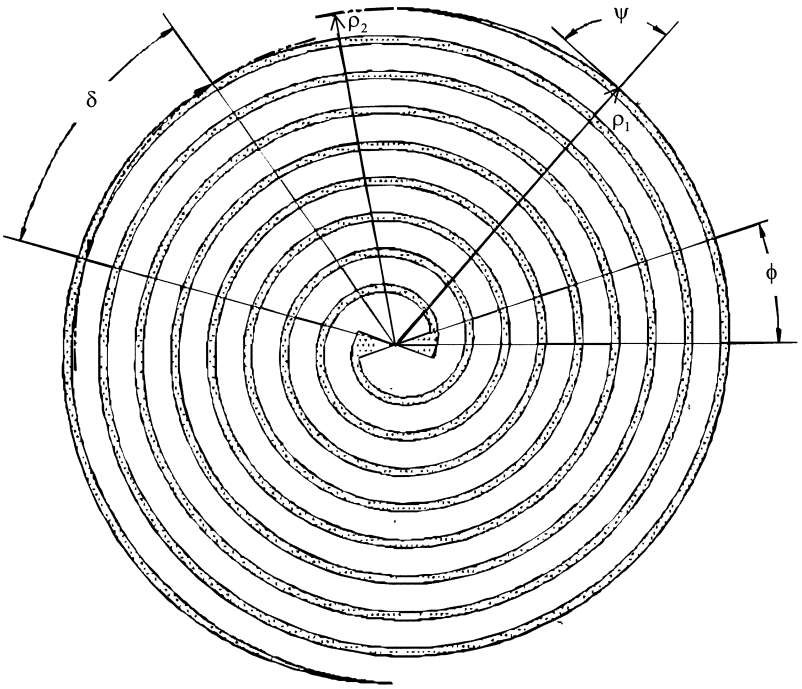


Figure 9.16 Planar two-arm Archimedean spiral antenna.

An example of an Archimedean spiral is shown in Fig. 9.16. The centerline of the antenna is defined by

$$\rho = a\,\phi \tag{9.45}$$

The first arm, which begins along the x-axis in this example, is defined by the edges

$$\begin{aligned} &\text{Outer edge:} \quad \rho_1 = a\left(\phi + \frac{\delta}{2}\right) \\ &\text{Inner edge:} \quad \rho_2 = a\left(\phi - \frac{\delta}{2}\right) \end{aligned} \tag{9.46}$$

where the pitch of the spiral arms is given by

$$\tan\psi = \frac{\rho}{a} \tag{9.47}$$

Again, additional arms are realized by successively rotating this arm by the angle $2\pi/N$. The arm width (W) and centerline spacing between adjacent arms (S) are given by

$$\begin{aligned} W &= a\delta\sin\psi \\ S &= 2\pi a\sin\psi \end{aligned} \tag{9.48}$$

© 2006 by Taylor & Francis Group, LLC

For a two-arm Archimedean, if $W/S = 1/4$, the structure is self-complementary. This is equivalent to $\delta = \pi/2$. The radiation properties of the Archimedean spiral are very similar to the log spiral.

As mentioned earlier, planar spiral antennas are bidirectional. To eliminate the back-side radiation, the antenna is usually backed by an absorber-filled cavity [27]. If appropriately designed, the cavity has little affect on the front-side pattern and impedance properties of the antenna, but it does reduce the efficiency of the structure to less than 50%. In addition, these cavities are rather deep. For low profile applications and to improve the efficiency of the antenna, printed spirals—spiral antennas printed on a relatively thin conductor-backed dielectric substrate—have been developed [28]. These printed spirals are typically Archimedean and are much narrower band than their cavity-backed counterparts, with typical bandwidths on the order of 2:1. The arms of these antennas essentially form microstrip lines; hence, the currents are more tightly bound to the structure than in other spiral designs. As a result, care must be taken to appropriately load the outer portions of the arms of the printed spirals in order to attenuate residual currents that propagate through the active region of the antenna [29]. In addition, the antenna cannot be made too large because of perturbing radiation from higher order modes, i.e., modes with active regions of circumference $(M + pN)\lambda$, where $p = 1, 2, 3, \ldots$ These higher order modes are excited by the residual currents. This size requirement also places limits on the operating bandwidth of the antenna.

Another technique used to realize unidirectional patterns is to place the equiangular spiral on a conical form, as shown in Fig. 9.17 [30]. The result, a conical equiangular spiral antenna, has a broad, single-lobed, circularly polarized pattern along the direction of the apex of the cone. On the conical form, with half angle θ_0, the equiangular spiral curve is defined as

$$r = r_i e^{(a \sin\theta_0)(\phi - \phi_i)} \qquad \text{for } \theta = \theta_0 \tag{9.49}$$

As shown in the figure, the flare rate is related to the pitch angle of the spiral by Eq. (9.41). The first arm is defined by the edges

$$\begin{aligned} &\text{Outer edge:} && r_1 = r_0 e^{(a \sin\theta_0)\phi} \\ &\text{Inner edge:} && r_2 = r_0 e^{(a \sin\theta_0)(\phi - \delta)} \end{aligned} \tag{9.50}$$

As before, a two-arm conical spiral is self-complementary when $\delta = \pi/2$. It is useful to note that the front-to-back ratio of the unidirectional pattern increases with pitch angle and decreases with increasing cone angle (θ_0) [31].

9.2.5. Log-periodic Antennas

If a particular antenna structure has the property that it is equal to itself after being increased in scale by a factor $1/\tau$, then the antenna will have the same electrical properties at the frequencies f and τf. As such,

$$\frac{f_L}{f_H} = \tau \qquad f_L < f_H \tag{9.51}$$

© 2006 by Taylor & Francis Group, LLC

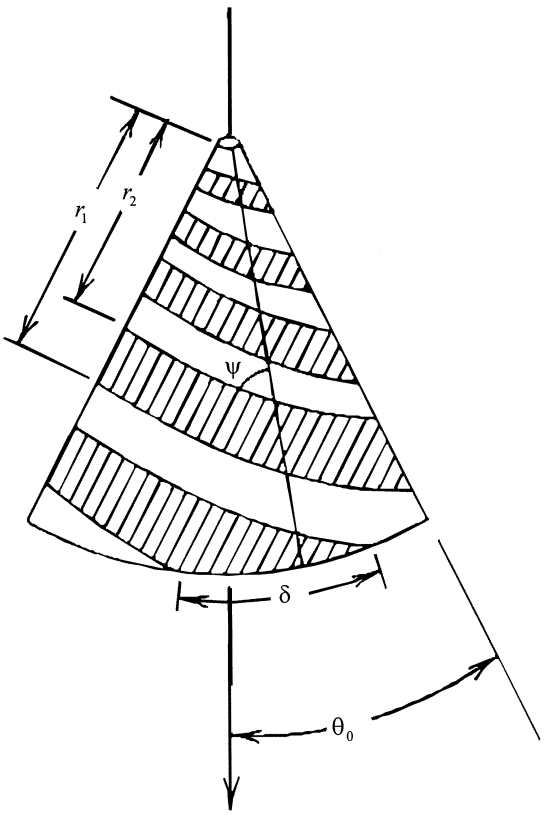


Figure 9.17 Conical two-arm equiangular spiral antenna.

Taking the logarithm of both sides of (9.51),

$$\log f_H = \log f_L + \log \frac{1}{\tau} \tag{9.52}$$

Hence, the electrical properties of the antenna are periodic when plotted on a log-frequency scale, with a period of $\log(1/\tau)$. Antennas that are based upon this principle are called *log-periodic* antennas. If these properties are relatively constant over the range $(f, \tau f)$, then they will be relatively constant for all frequencies, and the antenna will be quasi-frequency independent. Note that the equiangular spiral antenna is technically a log-periodic antenna (thus, its commonly referred to as a *log-spiral*) since the structure is unchanged when scaled by Eq. (9.44). The only variation in the electrical properties for an infinite log-spiral over the range $(f, \tau f)$ is a rotation of the radiation pattern.

The number of log-periodic antenna designs in use is too large to cover in this discussion [26]. Rather the general operating principles of a common structure will be given. Consider the planar trapezoidal-tooth log-periodic antenna shown in Fig. 9.18. The two elements are rotated versions of each other and connected to the feed at their respective vertex. The antenna is balanced; hence, a balanced feed must be used, usually with an impedance matching transformer. The number of teeth on each side of the center

© 2006 by Taylor & Francis Group, LLC

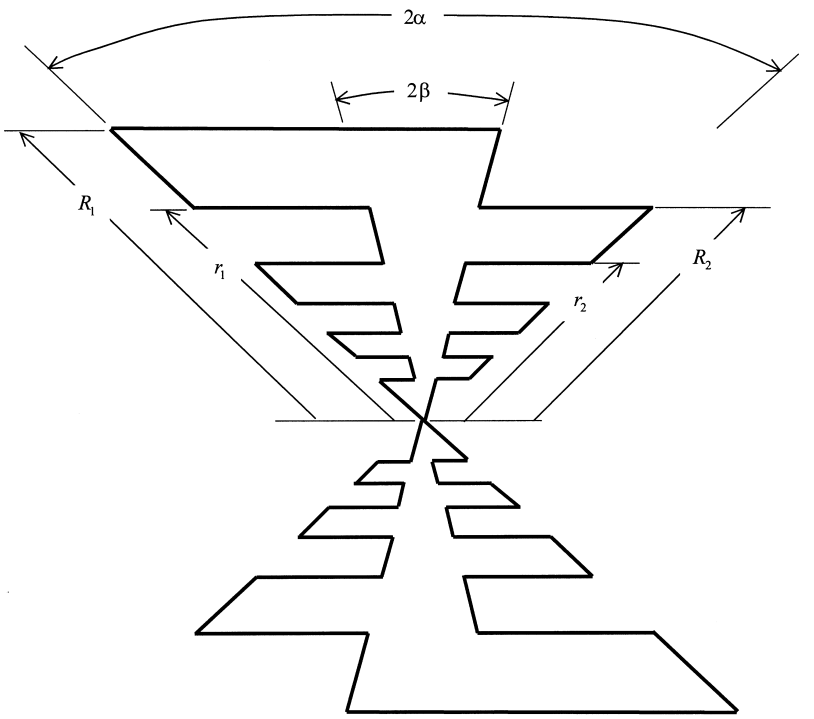


Figure 9.18 Planar trapezoidal-tooth log-periodic antenna.

strip (angular dimension 2β) should be the same and stagger spaced, as shown in the figure. Defining R_n as the distance from the vertex to the outer edge of the nth tooth ($n = 1$ for outer tooth, even n on one side of the strip, odd n on the other) and r_n as the distance from the vertex to the inner edge of the nth tooth, the scaling ratio for the structure is defined as

$$\tau = \frac{R_{n+2}}{R_n} < 1 \tag{9.53}$$

and the width of each tooth is characterized by

$$\varepsilon = \frac{r_n}{R_n} < 1 \tag{9.54}$$

Typically, $r_n = R_{n+1}$. With respect to the logarithm of frequency, the input impedance has a period $\log(1/\tau)$; however, given the staggered teeth and rotational symmetry of the arms, the pattern has a period $2\log(1/\tau)$.

For this structure, the angular center strips act as a transmission line, carrying current from the feed to the effective dipoles formed by the opposing teeth. Most of the radiation from the antenna occurs in the region where the effective dipole length on each arm is approximately $\lambda_0/2$. This is the active region of the antenna. The radiation

© 2006 by Taylor & Francis Group, LLC

pattern is bidirectional for the planar structure, essentially that of two-parallel dipoles, and it is polarized along the direction of the teeth. Unidirectional log-periodic antennas can be formed by bending the planar elements at the feed to form a wedge, similar in concept to the conical spiral. The low-frequency limit of the antenna occurs when the longest tooth is approximately $\lambda_0/4$. The high-frequency limit is established by the shortest teeth.

9.3. TRAVELING AND LEAKY-WAVE ANTENNAS

9.3.1. Introduction

Traveling-wave antennas are a class of antennas that use a traveling wave on a guiding structure as the main radiating mechanism [32,33]. They possess the advantage of simplicity, as no complicated feed network is required as in, for example, a typical array antenna.

Traveling-wave antennas fall into two general categories, slow-wave antennas and fast-wave antennas, which are usually referred to as *leaky-wave antennas* [34,35]. In a slow-wave antenna, the guided wave is a slow wave, meaning a wave that propagates with a phase velocity that is less than the speed of light in free space. Such a wave does not fundamentally radiate by its nature, and radiation occurs only at discontinuities (typically the feed and the termination regions). The propagation wave number of the traveling wave is therefore a real number (ignoring conductor or other losses). Because the wave radiates only at discontinuities, the radiation pattern physically arises from two equivalent sources, one at the beginning and one at the end of the structure. This makes it difficult to obtain highly directive single-beam radiation patterns. However, moderately direct patterns having a main beam near end fire can be achieved, although with a significant side-lobe level. For these antennas there is an optimum length depending on the desired location of the main beam.

By contrast, the wave on a leaky-wave antenna is a fast wave, with a phase velocity greater than the speed of light. This type of wave radiates continuously along its length, and hence the propagation wave number is complex [36,37], consisting of both a phase and an attenuation constant. Highly directive beams at an arbitrary specified angle can be achieved with this type of antenna, with a low side-lobe level. The phase constant of the-wave controls the beam angle, while the attenuation constant controls the beam width. The aperture distribution can also be easily tapered to control the side-lobe level or beam shape.

Leaky-wave antennas can be divided into two important categories, uniform and periodic, depending on the type of guiding structure [34]. A uniform structure has a cross section that is uniform (constant) along the length of the structure, usually in the form of a waveguide that has been partially opened to allow radiation to occur. The guided wave on the uniform structure is a fast wave, and thus radiates as it propagates.

A periodic leaky-wave antenna structure is one that consists of a uniform structure that supports a slow (nonradiating) wave that has been periodically modulated in some fashion. Since a slow wave radiates at discontinuities, the periodic modulations (discontinuities) cause the wave to radiate continuously along the length of the structure. From a more sophisticated point of view, the periodic modulation creates a guided wave that consists of an infinite number of space harmonics (Floquet modes). Although the main ($n=0$) space harmonic is a slow wave, one of the space harmonics (usually the $n=-1$) is designed to be a fast wave and hence a radiating wave.

© 2006 by Taylor & Francis Group, LLC

9.3.2. Slow-wave Antennas

A variety of guiding structures can be used to support a slow wave. Examples include wires in free space or over a ground plane, helixes, dielectric slabs or rods, corrugated conductors, etc. Many of the basic principles of slow-wave antennas can be illustrated by considering the case of a long wire antenna, shown in Fig. 9.19. The current on the wire is taken as

$$I(z') = I_0 e^{-j\beta z'} \qquad 0 < z' < L \tag{9.55}$$

where it is assumed that $\beta \geq k_0$. The far-field pattern is equal to the pattern of an infinitesimal unit amplitude electric dipole in the z direction multiplied by an array factor term, which is expressed in terms of the Fourier transform of the wire current. The far-zone electric field is polarized in the θ direction and is given by

$$E_\theta = \left(\frac{j\omega\mu_0}{4\pi r} e^{-jk_0 r} \sin\theta\right) \int_0^L I(z') e^{jk_0 z' \cos\theta} dz' \tag{9.56}$$

Substituting the current expression, Eq. (9.55), and performing the integral yields

$$E_\theta = \left(\frac{j\omega\mu_0}{4\pi r} e^{-jk_0 r} \sin\theta\right) I_0 \left[\frac{e^{jL(k_0\cos\theta - \beta)}}{j(k_0\cos\theta - \beta)} - \frac{1}{j(k_0\cos\theta - \beta)}\right] \tag{9.57}$$

Equation (9.56) indicates that the radiation comes (by superposition) from the entire length of the radiating current. The form of Eq. (9.57), on the other hand, makes it clear that the radiation may also be interpreted as coming from the two ends of the current segment. Both points of view are correct, since Eqs. (9.56) and (9.57) are mathematically equivalent. For calculation purposes, it is convenient to write Eq. (9.57) as

$$E_\theta = \left(\frac{j\omega\mu_0}{4\pi r} e^{-jk_0 r} \sin\theta\right) (I_0 L) e^{-j(L/2)(\beta/k_0 - \cos\theta)} \operatorname{sinc}\left[\frac{k_0 L}{2}\left(\frac{\beta}{k_0} - \cos\theta\right)\right] \tag{9.58}$$

where $\operatorname{sinc} x \equiv (\sin x)/x$.

The $\operatorname{sinc} x$ function is maximum at $x = 0$, which corresponds to an angle θ_0 in "invisible space," since $\cos\theta_0 = \beta/k_0 > 1$. This explains why it is not possible to obtain a narrow single-beam pattern with this type of current. Although the array factor (the sinc term) is maximum at end fire ($\theta = 0$), the presence of the $\sin\theta$ term from the element

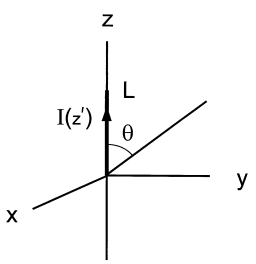
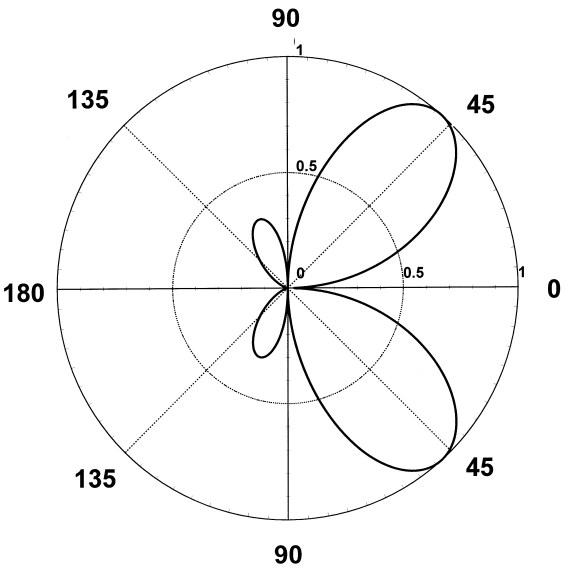


Figure 9.19 A traveling-wave current $I(z')$ on a wire, existing from $z = 0$ to $z = L$.

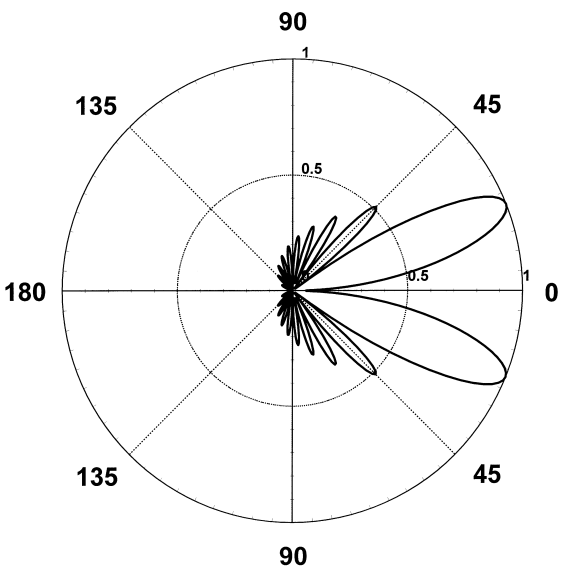
© 2006 by Taylor & Francis Group, LLC

pattern of the infinitesimal dipole results in a main beam that has a maximum shifted away from end fire.

Figure 9.20 shows patterns for the practical case $\beta/k_0 = 1$, for several lengths of current. It is seen that a longer current results in a "main beam" with a maximum closer to end fire, which is moderately more directive. However, the number of lobes in the pattern also increases with increasing current length. An independent control of the beam angle



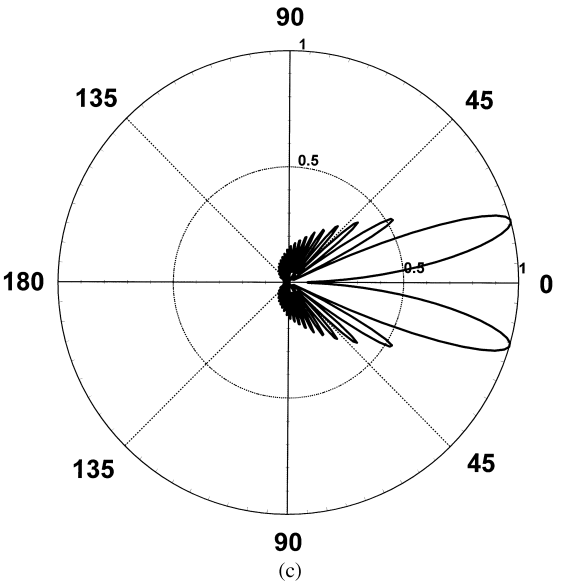
(a)



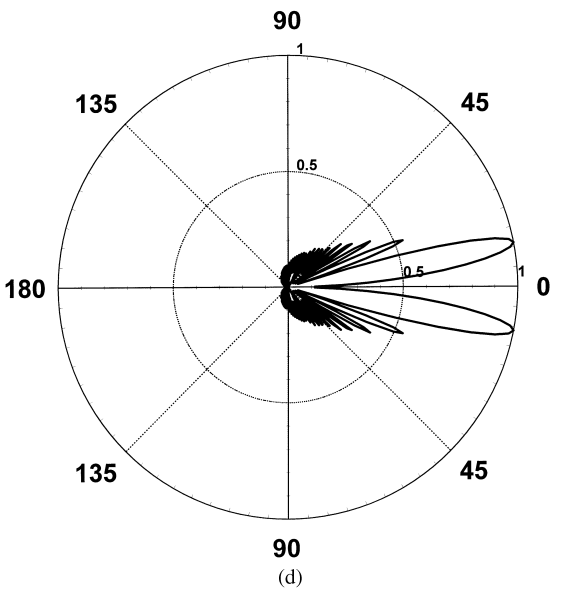
(b)

Figure 9.20 Far-field radiation patterns showing $E_\theta(\theta)$ for the traveling-wave current in Fig. 9.19, with $\beta = k_0$: (a) $L = 1.0\lambda_0$, (b) $L = 5.0\lambda_0$, (c) $L = 10.0\lambda_0$, and (d) $L = 20.0\lambda_0$.

© 2006 by Taylor & Francis Group, LLC



(c)



(d)

Figure 9.20 Continued.

and the beam width is not possible. Hence, the antenna length must be chosen in accordance with the desired angle of maximum radiation. An approximate formulas for the angle of maximum radiation is

$$\theta_0 = \cos^{-1}\left(1 - \frac{0.371}{L/\lambda_0}\right) \tag{9.59}$$

© 2006 by Taylor & Francis Group, LLC

Subsequent maxima occur at $\theta = \theta_m\,(m = 1, 2, 3, \ldots)$ given by [38]

$$\theta_m = \cos^{-1}\left(1 - \frac{p_m}{L/\lambda_0}\right) \tag{9.60}$$

where $p_m = 1.465, 2.48, 3.485, 4.495, 5.5, \ldots,$ for $m = 1, 2, 3, \ldots$. Note that as m increases, p_m approaches $(2m + 1)/2$.

The angle at which the nth null away from end fire appears is [38]

$$\theta_n = \cos^{-1}\left(1 - \frac{n}{L/\lambda_0}\right) \tag{9.61}$$

A practical arrangement for producing a traveling wave current is the horizontal wire at a height h over the earth (modeled as a perfect conductor), shown in Fig. 9.21. The wire and earth form (via image theory) a two-wire (twin-line) transmission line with separation $2h$ between the wires. A termination of the wire to the earth with a load resistance equal to twice the characteristic impedance Z_0 of the twin line will minimize reflections from the end. For this antenna the pattern of Eq. (9.58) is modified by multiplying by a factor $2j\sin(k_0 h\cos\theta)$. The pattern then becomes

$$E_\theta = \left(-\frac{j\omega\mu_0}{4\pi r}e^{-jk_0 r}\cos\theta\cos\phi\right)(I_0 L)e^{-j(L/2)(\beta/k_0 - \cos\theta_x)} \times 2j\sin(k_0 h\cos\theta)\mathrm{sinc}\left[\frac{k_0 L}{2}\left(\frac{\beta}{k_0} - \cos\theta_x\right)\right] \tag{9.62}$$

$$E_\phi = \left(\frac{j\omega\mu_0}{4\pi r}e^{-jk_0 r}\sin\phi\right)(I_0 L)e^{-j(L/2)(\beta/k_0 - \cos\theta_x)} \times 2j\sin(k_0 h\cos\theta)\mathrm{sinc}\left[\frac{k_0 L}{2}\left(\frac{\beta}{k_0} - \cos\theta_x\right)\right] \tag{9.63}$$

where $\cos\theta_x = \sin\theta\cos\phi$. Note that in the coordinate system of Fig. 9.21, the antenna radiates both polarizations. The electric field would be polarized in the θ direction if the coordinate system were rotated to align the z axis with the wire.

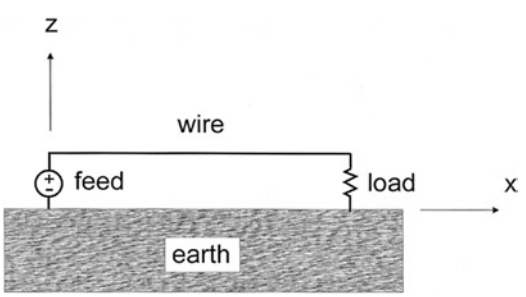


Figure 9.21 A practical traveling-wave antenna consisting of a long horizontal wire over the earth, terminated by a matched load resistor.

© 2006 by Taylor & Francis Group, LLC

9.3.3. Leaky-wave Antennas

As mentioned previously, a leaky-wave antenna (LWA) support a fast wave that radiates continuously along the length of the structure. The two types, uniform and periodic, are considered separately.

Uniform Structures

A typical example of a uniform leaky-wave antenna is a rectangular waveguide with a longitudinal slot, as shown in Fig. 9.22 [39]. Another variation on the design would be an array of closely spaced rectangular or circular slots in the waveguide wall instead of a long longitudinal slot [39]. Although, technically speaking, the periodic structure would not be a uniform structure, it could be modeled as such, provided the slots are closely spaced so that radiation comes only from the fundamental fast waveguide mode, and not a higher order Floquet mode (as for the periodic leaky-wave antennas discussed in the next section). The simple structure in Fig. 9.22 illustrates the basic properties common to all uniform leaky-wave antennas.

The fundamental TE_{10} waveguide mode is a fast wave, with $\beta < k_0$. In particular,

$$\frac{\beta}{k_0} = \sqrt{1 - \left(\frac{\pi}{k_0 a}\right)^2} \tag{9.64}$$

The fast-wave property of the aperture distribution in the slot causes the antenna to radiate similar to a phased array, with the angle of the beam from the waveguide axis given approximately by [37]

$$\cos\theta_0 \approx \frac{\beta}{k_0} \tag{9.65}$$

The radiation causes the wave number k_z of the propagating mode within the open waveguide structure to become complex, so that $k_z = \beta - j\alpha$. The constants β and α are

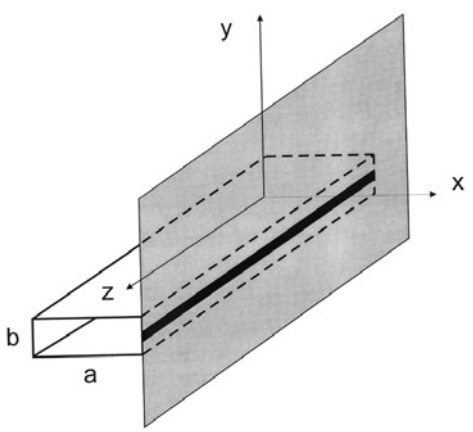


Figure 9.22 A leaky-wave antenna consisting of a rectangular waveguide with a long longitudinal slot in the narrow wall of the waveguide. An infinite ground plane surrounds the slot.

© 2006 by Taylor & Francis Group, LLC

referred to as the *phase* and *attenuation* constants, respectively. The phase constant controls the beam angle, and this can be varied by changing the frequency.

From image theory, the radiation from this structure in the region $x > 0$ is essentially due to a magnetic line current in free space of the form

$$K(z') = A\exp(-jk_z z') \tag{9.66}$$

that exists from $z' \in (0, \infty)$. The normalized pattern shape for the far-zone field $E_\phi(\theta)$ has the shape [37]

$$R(\theta) = \left|E_\phi(\theta)\right| = \left|\frac{\sin(\theta)}{(k_z/k_0) - \cos\theta}\right| \tag{9.67}$$

The attenuation constant controls the beam width of the pattern. An approximate formula for the beamwidth, measured between half-power points, is

$$BW = 2\csc\theta_0 \frac{\alpha}{k_0} \tag{9.68}$$

As is typical for a uniform LWA, the beam cannot be scanned too close to broadside ($\theta_0 = 90^\circ$), since this corresponds to the cutoff frequency of the waveguide. In addition, the beam cannot be scanned too close to end fire ($\theta_0 = 0^\circ$) since this requires operation at frequencies significantly above cutoff, where higher-order modes can propagate, at least for an air-filled waveguide. The $\sin\theta$ term in Eq. (9.67) also limits the end-fire scan. Scanning is limited to the forward quadrant only ($0 < \theta_0 < \pi/2$), for a wave traveling in the positive z direction.

This one-dimensional (1D) leaky-wave aperture distribution results in a "fan beam" having a narrow beam in the H plane (xz plane) with a beam width given by Eq. (9.67), and a broad beam in the E plane (xy plane). A pencil beam can be created by using an array of such 1D radiators.

H-plane patterns for the case $\beta/k_0 = 0.7071$ and $\alpha/k_0 = 0.1$ and 0.01 are shown in Fig. 9.23. This particular value of β corresponds to a beam angle of 45°. It is seen that, in accordance with Eq. (9.68), the pattern corresponding to the smaller α value has a much smaller beamwidth. Unlike the slow-wave structure, a very narrow beam can be created at any angle by choosing a sufficiently small value of α.

One interesting property of the leaky-wave antenna is the exponentially growing or "improper" nature of the near field surrounding the aperture region [36]. To understand this, consider an infinite line source that extends over the entire z axis, having the form of Eq. (9.66). In the air region surrounding the line source, the electric vector potential would have the form [40]

$$F_z = A\frac{\varepsilon_0}{4j}H_0^{(2)}(k_\rho\rho)e^{-jk_z z} \tag{9.69}$$

where

$$k_\rho = \left(k_0^2 - k_z^2\right)^{1/2} \tag{9.70}$$

© 2006 by Taylor & Francis Group, LLC

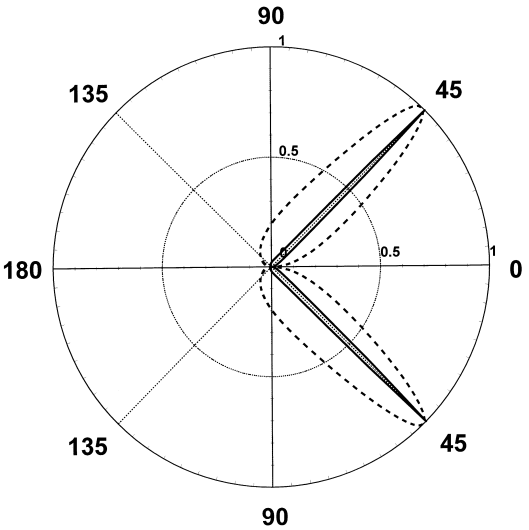


Figure 9.23 H-plane patterns for a leaky-wave line source of magnetic current existing in the semi-infinite region $0 < z < \infty$. The phase constant of the current wave is $\beta/k_0 = 0.7071$. Results are shown for two different values of the attenuation constant, $\alpha/k_0 = 0.1$ (dashed line) and 0.01 (solid line).

To further simplify this expression, consider large radial distances ρ from the z axis, so that the Hankel function may be asymptotically approximated, yielding

$$F_z = A\frac{\varepsilon_0}{4j}\sqrt{\frac{2j}{\pi k_\rho \rho}}e^{-jk_\rho \rho}e^{-jk_z z} \tag{9.71}$$

The wavenumber k_z is in the fourth quadrant of the complex plane. Therefore the radial wavenumber k_ρ from Eq. (9.70) must lie within either the first or third quadrants. Assuming that $\beta < k_0$, the physical choice is the one for which $\mathrm{Re}(k_\rho) > 0$, corresponding to an outward radiating wave. Hence k_ρ is within the first quadrant, and therefore $\mathrm{Im}(k_\rho) > 0$. That is, the wave field exponentially grows with radial distance away from the axis. For a leaky wave existing over the entire z axis, the radiation condition at infinity would be violated. However, for the semi-infinite line source existing over the region $(0, \infty)$ (corresponding to a leaky-wave antenna with a practical feed), the field surrounding the source grows only within an angular region defined by the leakage angle, as shown in Fig. 9.24 [36]. Outside this region the field decays rapidly. (In Fig. 9.24 the strength of the field is indicated by the closeness of the radiation arrows.)

A control of the beam shape may be achieved by tapering the slot width, so that the slot width w, and hence the attenuation constant α, is now a function of z. Suppose that it is desired to achieve a amplitude taper $A(z)$ in the line source amplitude. Approximately, the power radiated per unit length $P_L(z)$ is proportional to $|A(z)|^2$. The attenuation constant is related to $P_L(z)$ and to the power $P(z)$ flowing down the waveguide as [40]

$$\alpha(z) = \frac{P_L(z)}{2P(z)} = -\frac{1}{2P(z)}\frac{dP(z)}{dz} \tag{9.72}$$

© 2006 by Taylor & Francis Group, LLC

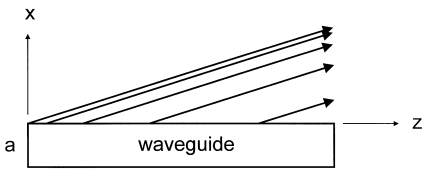


Figure 9.24 An illustration of the near-field behavior of a leaky wave on a guiding structure that begins at $z = 0$ (illustrated for the leaky-wave antenna of Fig. 9.22). The rays indicate the direction of power flow in the leaky-wave field, and the closeness of the rays indicates the field amplitude. This figure illustrates the exponential growth of the leaky-wave field in the x direction out to the leakage boundary.

Consider a finite length of radiating aperture, extending from $z = 0$ to $z = L$, with a terminating load at $z = L$ that absorbs all remaining power. After some manipulations, the formula for α can be cast into a form involving the desired aperture function $A(z)$ and the radiation efficiency e_r, defined as the power radiated divided by the total input power (the radiation efficiency is less than unity because of the load at the end). The result is [32]

$$\alpha(z) = \frac{(1/2)A^2(z)}{(1/e_r)\int_0^L A^2(z)dz - \int_0^Z A^2(z)dz} \tag{9.73}$$

A typical design would call for a 90% radiation efficiency ($e_r = 0.9$).

Equation (9.73) implies that the attenuation constant must become larger near the output (termination) end of the structure, and hence the loading (e.g., slot width) must become larger. In a practical design the loading would typically also be tapered to zero at the input (feed) end to ensure a gradual transition from the nonleaky to the leaky section of waveguide.

Periodic Structures

This type of leaky-wave antenna consists of a fundamentally slow-wave structure that has been modified by periodically modulating the structure in some fashion. A typical example is a rectangular waveguide that is loaded with a dielectric material and then modulated with a periodic set of slots, as shown in Fig. 9.25. Many of the features common to periodic leaky-wave antennas may be discussed by consideration of this simple structure.

It is assumed that the relative permittivity of the filling material is sufficiently high so that the TE_{10} mode is a slow wave over the frequency region of interest. This will be the case provided

$$\varepsilon_r > 1 + \left(\frac{\pi}{k_0 a}\right)^2 \tag{9.74}$$

for all values of k_0 in the range of interest. The waveguide mode is thus a nonradiating slow wave. However, because of the periodicity, the modal field of the periodically loaded waveguide is now in the form of a Floquet mode expansion [33],

$$E(x, y, z) = f(x, y) \sum_{n=-\infty}^{\infty} A_n e^{-jk_{zn}z} \tag{9.75}$$

© 2006 by Taylor & Francis Group, LLC

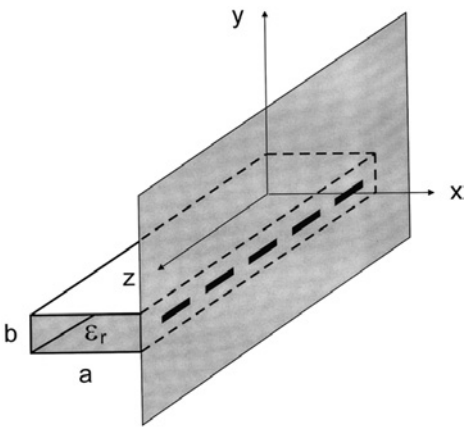


Figure 9.25 A periodic leaky-wave antenna consisting of a rectangular waveguide that is filled with a dielectric material and loaded with a periodic array of longitudinal slots in the narrow wall of the waveguide. An infinite ground plane surrounds the slots. The periodicity is d.

where

$$k_{zn} = k_{z0} + \frac{2\pi n}{d} \tag{9.76}$$

is the wave number of the nth Floquet mode or space harmonic. The zeroth wave number $k_{z0} = \beta - j\alpha$ is usually chosen to be the wave number that approaches the wave number of the closed waveguide when the loading (slot size) tends to zero. The wave number k_{z0} is then termed the propagation wave number of the guided wave.

Leakage (radiation per unit length of the structure) will occur provided one the space harmonics (usually the $n=-1$ space harmonic) is a fast wave, so that $-k_0 < \beta_{-1} < k_0$, where $\beta_{-1} = \beta - 2\pi/d$. By choosing the period d appropriately, the beam can be aimed from backward end fire to forward end fire. The beam will scan as the frequency changes, moving from backward end fire to forward end fire. If one wishes to have single-beam scanning over the entire range, the $n=-2$ space harmonic must remain a slow backward wave ($\beta_{-2} < -k_0$), while the fundamental space harmonic remains a slow forward wave ($\beta > k_0$) as the -1 space harmonic is scanned from backward to forward end fire. These design constraints result in the condition [41]

$$\varepsilon_r > 9 + \left(\frac{d}{a}\right)^2 \tag{9.77}$$

where a is the larger waveguide dimension.

One difficulty encountered in the scanning of periodic leaky-wave antennas is that the beam shape degrades as the beam is scanned through broadside. This is because the point $\beta_{-1} = 0$ corresponds to $\beta d = 2\pi$. This is a "stop band" of the periodic structure, where all reflections from the slot discontinuities add in phase back to the source [33]. At this point a perfect standing wave is set up within each unit cell of the structure, and the attenuation constant drops to zero. To understand this, consider the simple model of a transmission line (modeling the waveguide) periodically loaded with shunt loads

© 2006 by Taylor & Francis Group, LLC

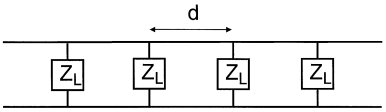


Figure 9.26 A simple approximate transmission line model for the periodic leaky-wave antenna in Fig. 9.7.

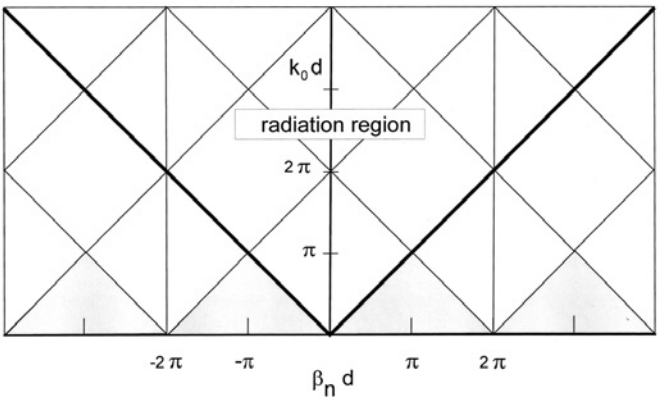


Figure 9.27 Brillouin, or $k-\beta$, diagram that is used for the physical explanation and interpretation of leakage from periodic structures. This diagram is a plot of k_0a versus βa, where a is the period and β is the phase constant of the guided mode (the phase constant of the $n=0$ space harmonic).

(modeling the slots), as shown in Fig. 9.26. When the electrical distance between the loads becomes one wavelength (corresponding to $\beta d = 2\pi$), the total input admittance at any load location becomes infinite (a short circuit). The power absorbed by the loads (radiated by the slots) therefore becomes zero. There are various ways in which the stop-band effect can be minimized. One method is to introduce two radiating elements per unit cell, spaced a distance $d/4$ apart within each cell [42]. At the stop-band point where $\beta d = 2\pi$, the electrical distance between the adjacent elements within the unit cell will be $\pi/2$. The round-trip phase delay between the two elements will then be 180°, which tends to minimize the effects of the reflection from the pair of elements.

When designing, analyzing, and interpreting periodic leaky-wave antennas, a useful tool is the $k-\beta$ or Brillouin diagram [33]. This is a plot of k_0d versus β_nd, as shown in Fig. 9.27. The darker lines on the diagram indicate boundaries where the $n=-1$ space harmonic will be radiating at backward end fire and forward end fire. The shaded regions (the regions inside the lower triangles) are the bound-wave regions [33]. For points in these regions, all of the space harmonics are slow (nonradiating) waves. For a point outside the bound-wave triangles, there must be at least one space harmonic that is a radiating fast wave.

Two-dimensional Leaky-wave Antennas

A broadside pencil beam, or a scanned conical beam, may be obtained by using a two-dimensional (2D) LWA, which supports a radially propagating cylindrical leaky wave instead of a 1D linearly propagating wave. One example of such a structure is the leaky

© 2006 by Taylor & Francis Group, LLC

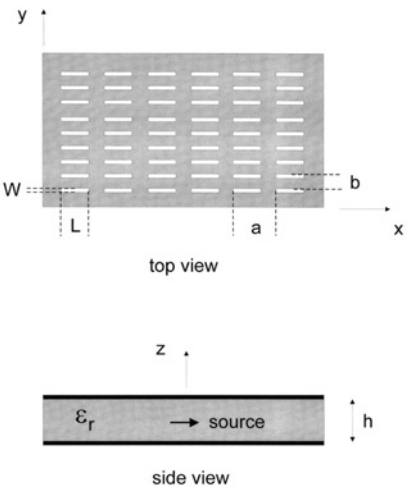


Figure 9.28 A two-dimensional leaky-wave antenna consisting of a periodic array of slots in a top plane, over a grounded dielectric slab. This structure acts as a leaky parallel-plate waveguide that is operating in the first higher order waveguide mode. The structure is excited by a simple source such as a horizontal electric dipole.

parallel-plate waveguide antenna shown in Fig. 9.28, which turns the first higher-order parallel-plate waveguide mode into a leaky mode by allowing radiation to occur through the slots [43]. (Although there is a periodic arrangement of slots, the structure is acting as a uniform leaky parallel-plate waveguide, due to the close spacing of the slots. Another design variation would use a high-permittivity dielectric layer instead of the slotted plate [44].) A simple source such as a horizontal dipole may be used to excite the radial leaky mode. The height h is chosen according to the desired beam angle θ_p. The radial waveguide mode is designed to be a fast wave with a phase constant

$$\beta = \left[k_0^2 \varepsilon_r - \left(\frac{\pi}{h}\right)^2\right]^{1/2} \tag{9.78}$$

The beam angle θ_p is measured from broadside and is related to the phase constant as $\beta = k_0 \sin\theta_p$. Solving for the plate separation yields

$$\frac{h}{\lambda_0} = \frac{0.5}{\sqrt{\varepsilon_r - \sin^2\theta_p}} \tag{9.79}$$

An example of a beam produced by such antenna is shown in Fig. 9.29 at a frequency of 12 GHz, using an air substrate. Patterns are shown for three different values of the substrate thickness h, demonstrating how the beam scans from broadside when the substrate thickness is increased. The three thicknesses chosen correspond to a scan angle of $\theta_p = 0°$, 15°, and 30°. The excitation is taken as a simple horizontal y-directed electric dipole in the middle of the air region, directly below the center slot. The structure is assumed to be infinite in the horizontal directions. Near the beam peak, a pencil beam is obtained with nearly equal beam widths in the E and H planes. Further details may be found in Ref. 43.

© 2006 by Taylor & Francis Group, LLC

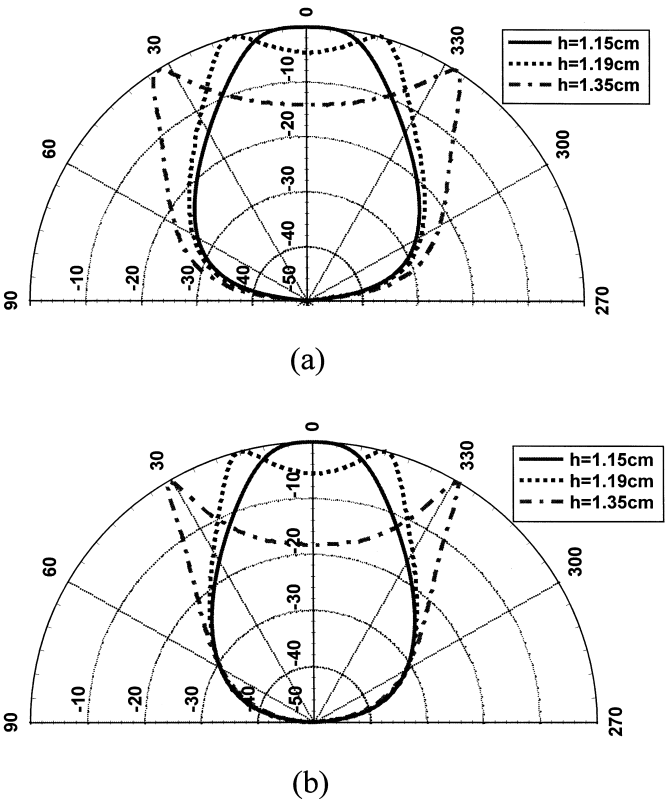


Figure 9.29 Radiation patterns for two-dimensional leaky-wave antenna shown in Fig. 9.28. Patterns are shown for three different values of the substrate thickness ($h = 1.15$ cm, 1.19 cm, 1.35 cm) to illustrate how the beam changes from a pencil beam at broadside to a scanned conical beam as the substrate thickness increases. An air substrate is assumed, and the frequency is 12 GHz: (a) E-plane patterns and (b) H-plane patterns. $l = 0.8$ cm, $w = 0.05$ cm, $a = 1.0$ cm, $b = 0.3$ cm. The y-directed source dipole is in the middle of the air substrate, directly below one of the slots.

9.4. APERTURE ANTENNAS

9.4.1. Introduction

There are classes of antennas that have a physical aperture through which the structure radiates electromagnetic energy. Examples are horn antennas, slotted-waveguide antennas, and open-ended waveguide. In addition, many antennas are more conveniently represented for analysis or qualitative understanding by equivalent apertures. In this section, a general analytical treatment of radiation from apertures is summarized. This approach is used to show the basic characteristics of some common aperture antennas. Also, a discussion of reflector antennas, in the context of an equivalent aperture, is presented.

9.4.2. Radiation from Apertures

Consider a general radiator, as shown in Fig. 9.30. Using the equivalence principle [40], the sources in the closed region S can be removed and equivalent surface-current densities

© 2006 by Taylor & Francis Group, LLC

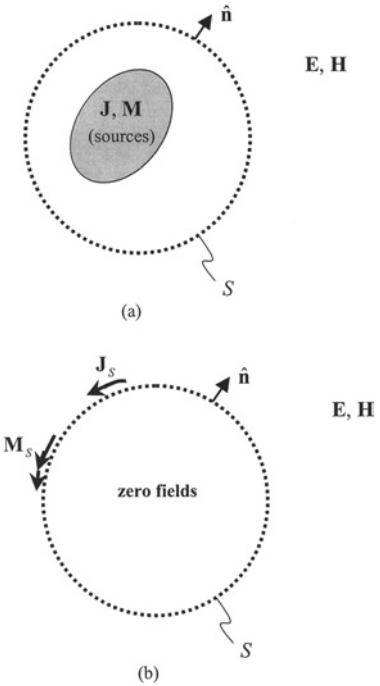


Figure 9.30 Equivalent aperture representation for a general radiator: (a) original problem and (b) equivalent problem.

($\mathbf{J}_S$, $\mathbf{M}_S$) placed on the equivalent aperture surface S. If the fields inside S are assumed to be zero, the equivalent current densities on S are given by

$$\mathbf{J}_S = \hat{\mathbf{n}} \times \mathbf{H}_A \qquad \mathbf{M}_S = \mathbf{E}_A \times \hat{\mathbf{n}} \tag{9.80}$$

where $\mathbf{E}_A$ and $\mathbf{H}_A$ are the fields on S produced by the original sources. These equivalent currents produce the same fields as the original sources in the region outside of S. In the far field the radiated electric field is given by

$$\mathbf{E}(\mathbf{J}_S, \mathbf{M}_S) \approx -j\omega - \hat{\mathbf{r}} \times \hat{\mathbf{r}} \times A - j\omega\sqrt{\mu\varepsilon}\,\mathbf{F} \times \hat{\mathbf{r}} \tag{9.81}$$

where only the θ and ϕ components are used. The far-field potential vectors are given by

$$\begin{aligned} \mathbf{A} &\approx \frac{e^{-jkr}}{4\pi r} \int_S \mathbf{J}_S(\hat{\mathbf{r}}') e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'} dS' \\ \mathbf{F} &\approx \frac{e^{-jkr}}{4\pi r} \int_S \mathbf{M}_S(\hat{\mathbf{r}}') e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'} dS' \end{aligned} \tag{9.82}$$

© 2006 by Taylor & Francis Group, LLC

Substituting Eq. (9.80) into Eq. (9.82) yields [25]

$$\begin{aligned}\mathbf{A} &\approx \frac{e^{-jkr}}{4\pi r}\hat{\mathbf{n}} \times \int_S \mathbf{H}_A(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}dS' = \frac{e^{-jkr}}{4\pi r}\hat{\mathbf{n}} \times \mathbf{Q} \\ \mathbf{F} &\approx -\frac{e^{-jkr}}{4\pi r}\hat{\mathbf{n}} \times \int_S \mathbf{E}_A(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}dS' = -\frac{e^{-jkr}}{4\pi r}\hat{\mathbf{n}} \times \mathbf{P}\end{aligned} \tag{9.83}$$

where

$$\begin{aligned}\mathbf{Q} &= \int_S \mathbf{H}_A(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}dS' \\ \mathbf{P} &= \int_S \mathbf{E}_A(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}dS'\end{aligned} \tag{9.84}$$

In many practical antenna problems, the radiating sources lie in a half space ($z < 0$) and an equivalent aperture surface can be defined in the xy plane ($\hat{\mathbf{n}} = \hat{\mathbf{z}}$). Hence, for this case Eqs. (9.81), (9.83), and (9.84) reduce to the following expressions for the radiated electric fields for $z > 0$:

$$\begin{aligned}E_\theta &\approx j\omega\mu\frac{e^{-jkr}}{4\pi r}\left[\sqrt{\frac{\varepsilon}{\mu}}\left(P_x\cos\phi + P_y\sin\phi\right) + \cos\theta\left(Q_y\cos\phi - Q_x\sin\phi\right)\right] \\ E_\phi &\approx j\omega\mu\frac{e^{-jkr}}{4\pi r}\left[\cos\theta\sqrt{\frac{\varepsilon}{\mu}}\left(P_y\cos\phi - P_x\sin\phi\right) - \left(Q_x\cos\phi + Q_y\sin\phi\right)\right]\end{aligned} \tag{9.85}$$

where

$$\begin{aligned}\mathbf{Q} &= \int_S \mathbf{H}_A(x', y')e^{jk(x'\sin\theta\cos\phi + y'\sin\theta\sin\phi)}dx'dy' = \int_S \mathbf{H}_A(\rho', \phi')e^{jk\rho'\sin\theta\cos(\phi-\phi')}\rho' d\rho' d\phi' \\ \mathbf{P} &= \int_S \mathbf{E}_A(x', y')e^{jk(x'\sin\theta\cos\phi + y'\sin\theta\sin\phi)}dx'dy' = \int_S \mathbf{E}_A(\rho', \phi')e^{jk\rho'\sin\theta\cos(\phi-\phi')}\rho' d\rho' d\phi'\end{aligned} \tag{9.86}$$

Notice that the integrals in Eq. (9.86) are simply two-dimensional Fourier transforms of the aperture fields.

For many planar aperture antennas, a physical aperture is cut into a conducting ground plane (xy plane). For this discussion, a physical aperture is a slot or hole in a conductor through which radiated electromagnetic waves emanate. The components of the electric field tangent to the ground plane are equal to zero except in the aperture; hence, $\mathbf{M}_S$ is nonzero only in the aperture. In general, $\mathbf{J}_S$ is nonzero over the entire xy plane. In addition, for many directive antennas, the fields in the xy plane are often approximated as zero except in the aperture (even when no ground plane is present). Since the fields in the $z < 0$ half space are assumed to be zero, it is usually convenient to replace this half space with a perfect electric conductor. As a result, using image theory, the equivalent $\mathbf{J}_S$ sources in the xy plane are shorted out and the equivalent $\mathbf{M}_S$ sources double in strength [40]. Therefore, in Eq. (9.86) $\mathbf{H}_A \to 0$, $\mathbf{E}_A \to 2\mathbf{E}_A$, and the integration over the entire xy

© 2006 by Taylor & Francis Group, LLC

plane reduces to integration only over the support of $\mathbf{M}_S$ (aperture surface S_A, where $\mathbf{E}_A$ is nonzero). The directivity of a planar aperture is given by [25]

$$D = \frac{4\pi}{\lambda^2} \frac{\left| \int_{S_A} \mathbf{E}_A \, dS' \right|^2}{\int_{S_A} |\mathbf{E}_A|^2 \, dS'} = \frac{4\pi}{\lambda^2} A_e \tag{9.87}$$

where A_e is the effective aperture area.

In the next few sections, this aperture analysis will be used to understand the basic radiation properties of the most common aperture-type antennas.

9.4.3. Electrically Small Rectangular Slot

Consider the rectangular slot aperture shown in Fig. 9.31, where it will be assumed that $L, W \ll \lambda_0$. A common approximation for small slots is to assume a uniform aperture electric field distribution

$$\mathbf{E}_A \approx \begin{cases} E_0 \hat{\mathbf{y}} & |x| \le L/2, |y| \le W/2 \\ 0 & \text{otherwise} \end{cases} \tag{9.88}$$

Using this in Eq. (9.85) yields the following expressions for the radiated fields:

$$E_\theta \approx j\omega\sqrt{\mu\varepsilon} \frac{e^{-jkr}}{2\pi r} E_0(WL) \sin\phi \left\{ \frac{\sin[(kW/2)\sin\theta\cos\phi]}{(kW/2)\sin\theta\cos\phi} \frac{\sin[(kL/2)\sin\theta\sin\phi]}{(kL/2)\sin\theta\sin\phi} \right\}$$

$$E_\phi \approx j\omega\sqrt{\mu\varepsilon} \frac{e^{-jkr}}{2\pi r} E_0(WL) \cos\theta\cos\phi \left\{ \frac{\sin[(kW/2)\sin\theta\cos\phi]}{(kW/2)\sin\theta\cos\phi} \frac{\sin[(kL/2)\sin\theta\sin\phi]}{(kL/2)\sin\theta\sin\phi} \right\} \tag{9.89}$$

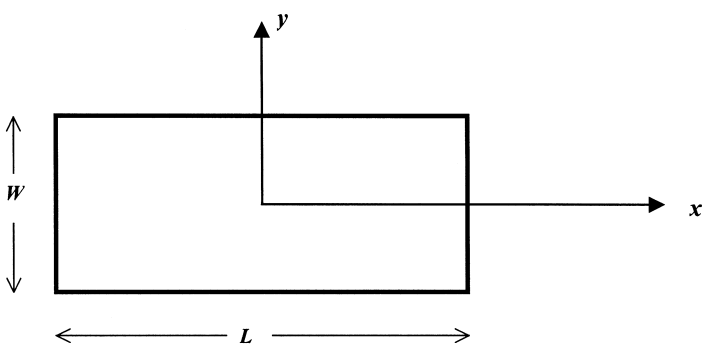


Figure 9.31 Electrically small rectangular slot aperture ($L, W \ll \lambda_0$).

© 2006 by Taylor & Francis Group, LLC

If L, $W \ll \lambda_0$, Eq. (9.89) reduces to

$$E_\theta \approx j\omega\sqrt{\mu\varepsilon}\frac{e^{-jkr}}{2\pi r}E_0 A_p \sin\phi$$
$$E_\phi \approx j\omega\sqrt{\mu\varepsilon}\frac{e^{-jkr}}{2\pi r}E_0 A_p \cos\theta\cos\phi \tag{9.90}$$

where $A_p = WL$ is the physical area of the aperture. For a uniform aperture field $A_e = A_p$; thus, the directivity is

$$D = \frac{4\pi}{\lambda^2}A_p \tag{9.91}$$

9.4.4. Rectangular Horn Antenna

Horn antennas are common high-frequency antennas for moderately high gain applications and applications where exact knowledge of the gain is required (theoretical calculations of the gain are very accurate). There are a number of different horn designs, including those with rectangular apertures (pyramidal, sectoral) and circular apertures (conical) [45,46]. Most use only a single waveguide mode to form the aperture distribution; however, multimode and hybrid-mode horns designs are also used for many specialized applications. In this section the focus will be single-mode horns with rectangular apertures and emphasis will be given to the pyramidal horns since they are the most commonly used horn designs.

Consider the pyramidal horn shown in Fig. 9.32. The horn is excited by the dominant TE_{10} mode of the feeding rectangular waveguide. The electric field distribution in the aperture of the antenna (xy plane) results from this mode propagating from the feed waveguide to the aperture of the horn. The aperture field appears to emanate from the TE_{10} fields at the apex of the horn. In the aperture, the transverse amplitude variation of TE_{10} electric field is preserved; however, the uniform phase (with respect to z) of the exciting mode is not maintained in the aperture since the wave has to propagate different distances to the various points in the aperture. This phase variation from the apex is given by [25]

$$e^{-jk_0(R-R_1)x}e^{-jk_0(R-R_2)y} \tag{9.92}$$

For relatively long horns, $A/2 \ll R_1$ and $B/2 \ll R_2$, the following approximations are commonly used:

$$R - R_1 \approx \frac{1}{2}\frac{x^2}{R_1}$$
$$R - R_2 \approx \frac{1}{2}\frac{y^2}{R_2} \tag{9.93}$$

© 2006 by Taylor & Francis Group, LLC

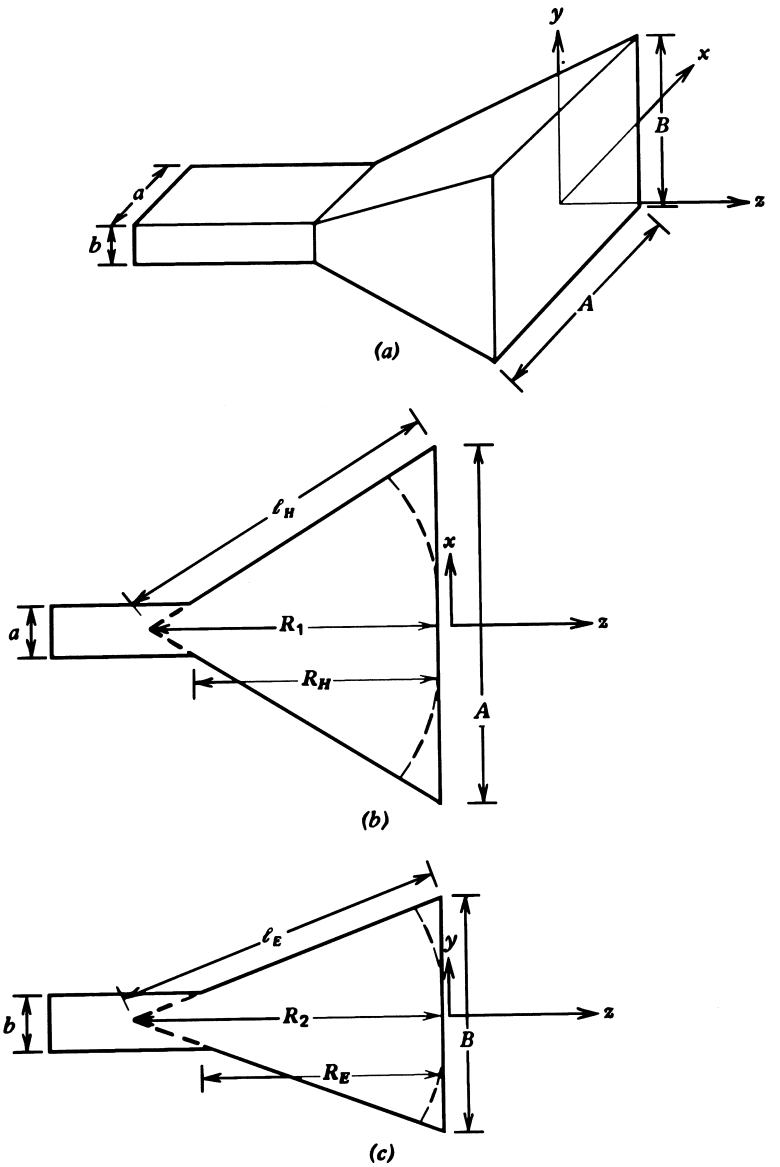


Figure 9.32 Pyramidal horn antenna: (a) perspective view, (b) H-plane (xz plane) cross section, and (c) E-plane (yz plane) cross section.

This leads to an aperture distribution given by

$$\mathbf{E}_A = \hat{\mathbf{y}} E_A = \hat{\mathbf{y}} \cos\frac{\pi x}{A} e^{-j(k_0/2R_1)x^2} e^{-j(k_0/2R_2)y^2} \tag{9.94}$$

Substituting this aperture distribution into Eq. (9.86) yields an integral that can be performed in closed form in the principle planes of the antenna; however, the results are

© 2006 by Taylor & Francis Group, LLC

rather complicated in form, involving Fresnel integrals [45]. Plots of the E-plane and H-plane patterns for various horn flares are shown in Fig. 9.33.

The directivity for a horn can be calculated using Eq. (9.87). If there is no phase variation across the aperture (idealized case) the effective aperture area is

$$A_e = \frac{8}{\pi^2} A_p \qquad \text{(uniform aperture phase)} \tag{9.95}$$

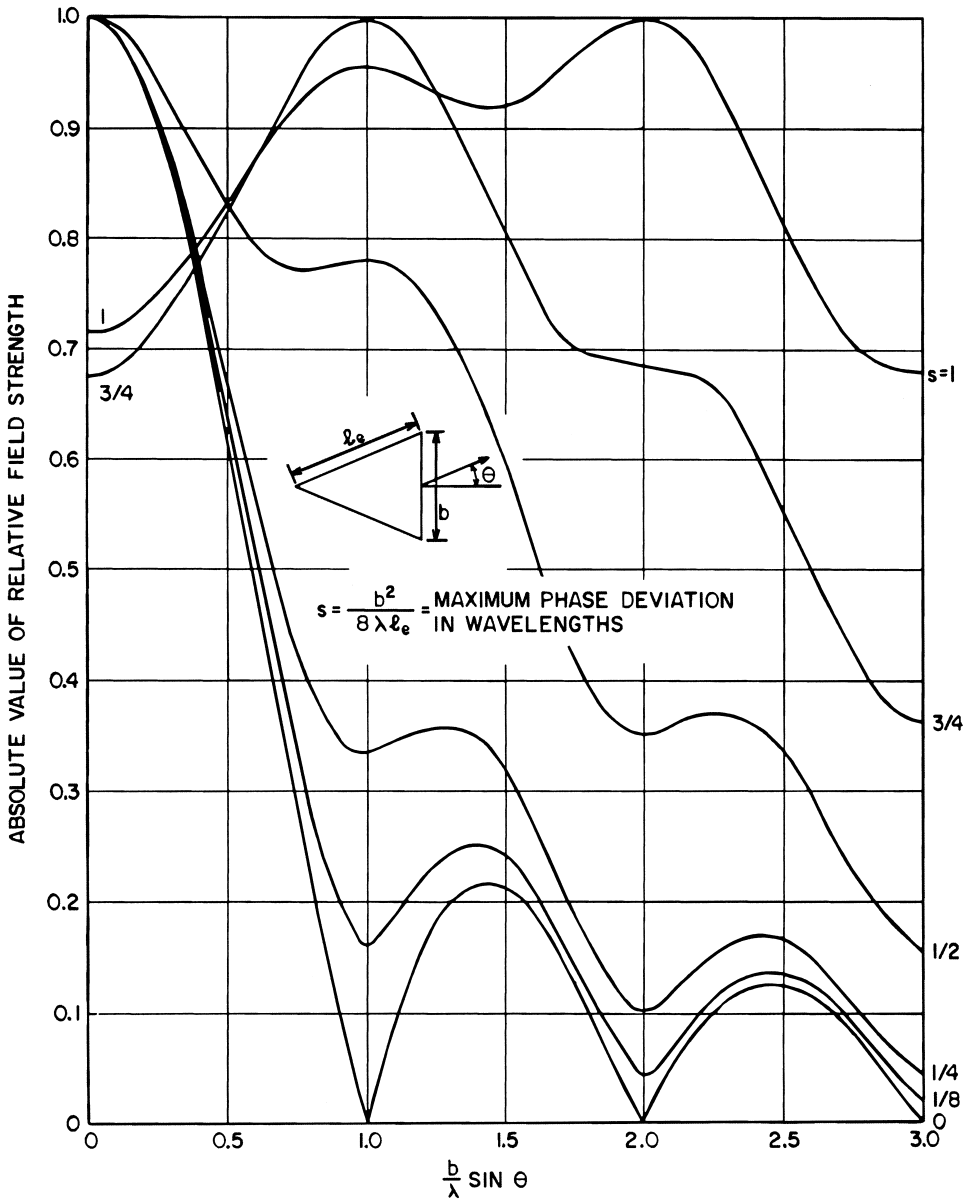
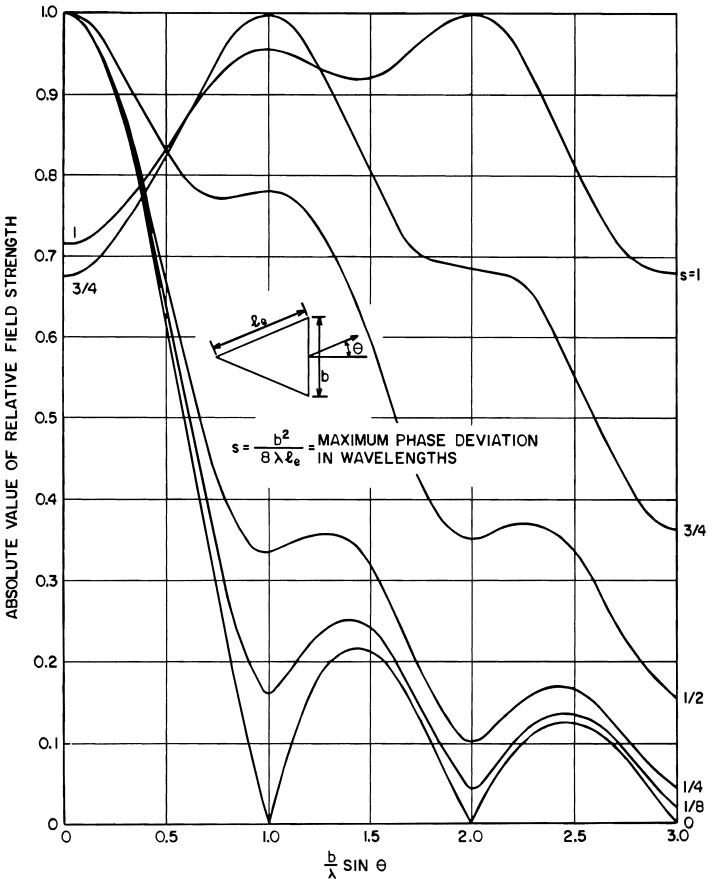


Figure 9.33 Universal radiation patterns for a rectangular horn antenna: (a) E plane and (b) H plane.

© 2006 by Taylor & Francis Group, LLC



(a)

Figure 9.33 Continued.

For an optimum pyramidal horn design, a design that produces the maximum directivity along boresight in the E and H planes ($A \approx \sqrt{3\lambda\ell_H}$ and $B \approx \sqrt{2\lambda\ell_E}$) [45], the effective aperture area is

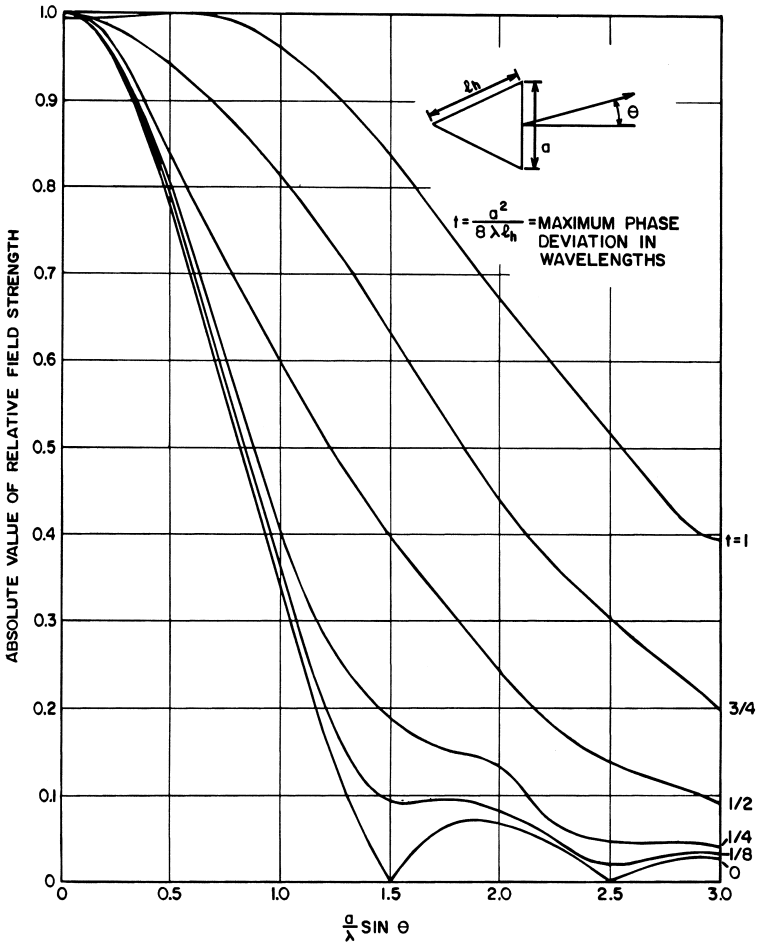
$$A_e = \frac{1}{2}A_p \qquad \text{(optimum pyramidal design)} \tag{9.96}$$

Typically, horn antennas have effective aperture areas that are 40–80% of the physical aperture. Another accurate approach to determine the directivity of a pyramidal horn is to use normalized directivity curves for E- and H-plane sectoral horns [25] as

$$D \approx \frac{\pi}{32}\left(\frac{\lambda}{A}D_E\right)\left(\frac{\lambda}{B}D_H\right) \tag{9.97}$$

where the normalized sectoral-horn directivities (terms in parenthesizes) are shown in Fig. 9.34.

© 2006 by Taylor & Francis Group, LLC



(b)

Figure 9.33 Continued.

9.4.5. Reflector Antennas

For high-gain antenna applications, applications requiring gains of 30 dB or more, reflector antennas are by far the most widely used. The design of these antennas is relatively complex, well beyond the scope of this discussion. In addition, the sheer number of reflector types is too numerous to summarize here [47,48]. However, it is very useful in understanding and designing reflector antennas to think of them in terms of aperture antennas, where the feed and reflector combination establish an aperture distribution that is radiated using the methods described earlier. The objective of this short discussion will therefore be limited to how to calculate an aperture electric field distribution for a simple parabolic reflector antenna.

Consider the parabolic reflector antenna shown in Fig. 9.35. The parabolic reflector is shaped such that the lengths of all ray paths from the feed to the reflector and then to the aperture plane (xy plane) are equal to twice the focal length ($2f$). As such, if the phase of the radiation pattern for the feed antenna is a constant, then the phase distribution in the

© 2006 by Taylor & Francis Group, LLC

aperture plane will be a constant. The description for the parabolic surface of the reflector is given by [25]

$$r' = f \sec^2 \frac{\theta'}{2} = \frac{4f^2 + \rho'^2}{4f} \tag{9.98}$$

The displacement from the focal point to any point in the aperture plane is

$$\rho' = r' \sin\theta' = 2f \tan\frac{\theta'}{2} \tag{9.99}$$

Using a geometrical optics or ray argument, it can be readily demonstrated that the amplitude distribution in the aperture plane is a function of the radiation pattern of the

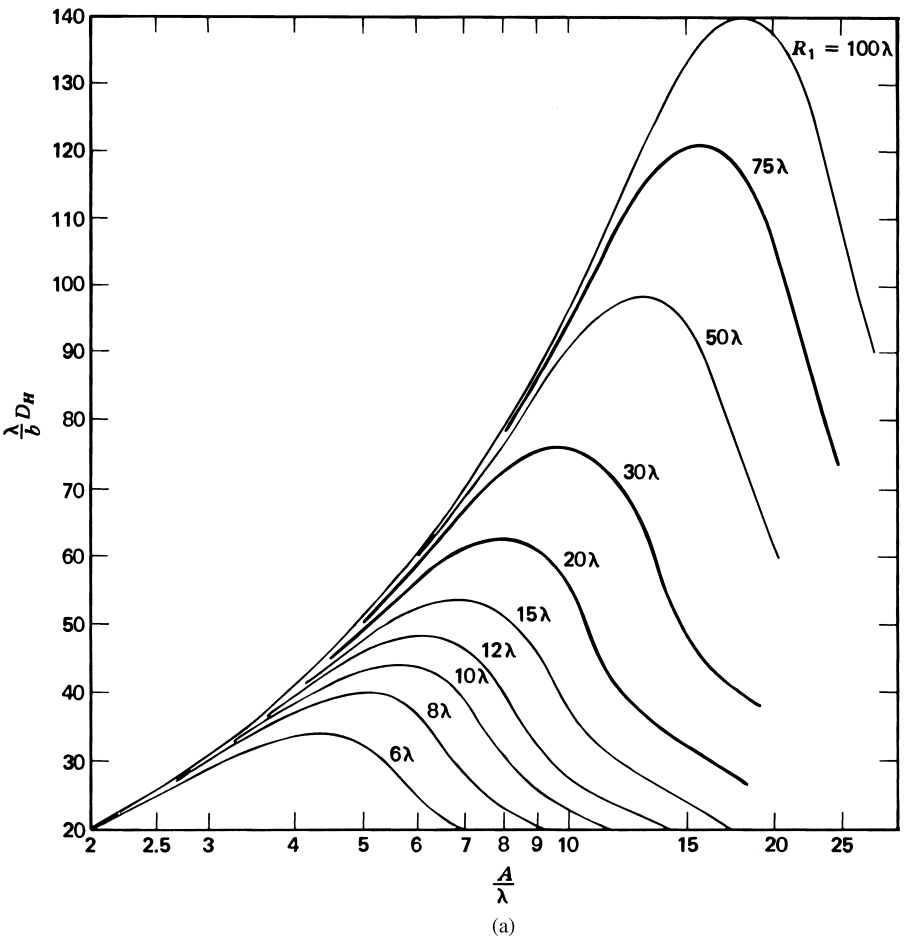


Figure 9.34 Universal directivity curves for rectangular sectoral horn antennas: (a) E plane and (b) H plane.

© 2006 by Taylor & Francis Group, LLC

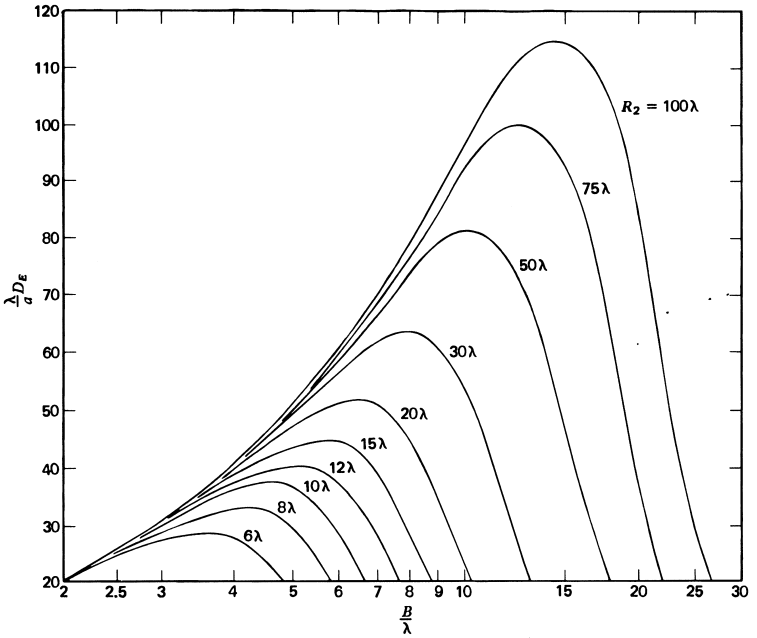


Figure 9.34 Continued.

feed antenna as

$$\mathbf{E}_A(\theta', \phi') = E_0 \frac{F(\theta', \phi')}{r'} \hat{\mathbf{l}}_A \tag{9.100}$$

where $F(\theta', \phi')$ is the normalized pattern of the feed antenna, $\hat{\mathbf{l}}_A$ is a unit vector in the direction of the aperture electric field given by

$$\hat{\mathbf{l}}_A = 2(\hat{\mathbf{n}} \cdot \hat{\mathbf{l}}_F)\hat{\mathbf{n}} - \hat{\mathbf{l}}_F \tag{9.101}$$

$\hat{\mathbf{l}}_F$ is a unit vector in the direction of the electric field radiated by the feed antenna, and $\hat{\mathbf{n}}$ is the unit normal to the surface of the reflector. Finally, the radiation field from the reflector antenna is determined by substituting Eqs. (9.98)–(9.101) into Eqs. (9.85) and (9.86).

9.5. PHASED ARRAYS

9.5.1. Array Far Fields

Phased arrays are arrays of antenna elements for which a radiation beam may be scanned electronically. We first examine the far-field pattern, i.e., the beam characteristics of the array. Consider a planar array of elements uniformly spaced in the $z = 0$ plane. The element locations may be defined via lattice vectors $\mathbf{d}_1$ and $\mathbf{d}_2$, as shown in Fig. 9.36a. Scanning of the array pattern is accomplished by introducing a constant progressive phase

© 2006 by Taylor & Francis Group, LLC

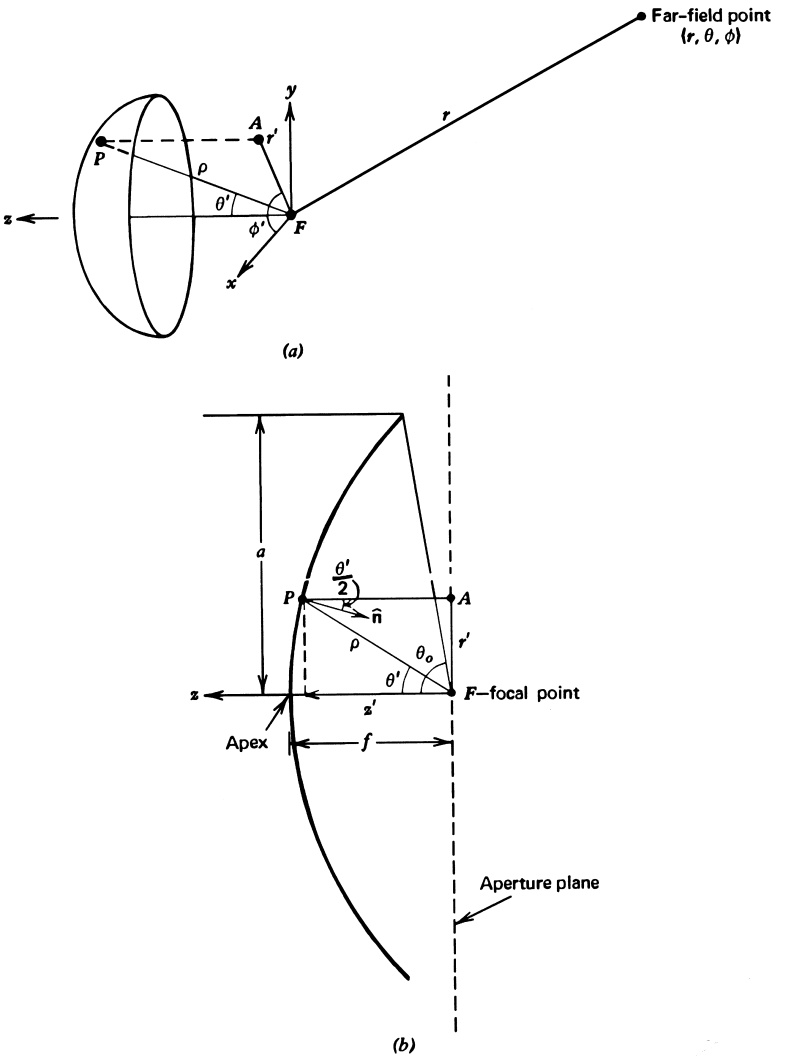


Figure 9.35 Parabolic reflector antenna: (a) perspective view and (b) cross-sectional view.

shift between elements. To determine the radiation pattern, we make the usual assumption, valid for arrays of a large number of elements, that the induced or equivalent current on the (m, n)th array element, $\mathbf{J}_{mn}(\mathbf{r})$, is identical to that of the $(0, 0)$th reference element, $\mathbf{J}_{0,0}(\mathbf{r})$, except for a positive, real amplitude factor a_{mn} and a phase shift ϕ_{mn}:

$$\mathbf{J}_{mn}(\mathbf{r} + m\mathbf{d}_1 + n\mathbf{d}_2) = a_{mn}\mathbf{J}_{0,0}(\mathbf{r})e^{j\phi_{mn}} \tag{9.102}$$

Hence the vector potential in the far field is given by the superposition

$$\begin{aligned}\mathbf{A}(\mathbf{r}) &= \frac{\mu_0}{4\pi r}e^{-jkr}\left(\sum_{m,n} a_{mn}e^{j\left[k\hat{\mathbf{r}}\cdot(m\mathbf{d}_1+n\mathbf{d}_2)+\phi_{mn}\right]}\right)\int_{S_{\text{ref}}}\mathbf{J}_{0,0}(\mathbf{r}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}\,dS' \\ &= \mathbf{A}_{\text{ref}}(\mathbf{r})\text{AF}(\hat{\mathbf{r}})\end{aligned} \tag{9.103}$$

© 2006 by Taylor & Francis Group, LLC

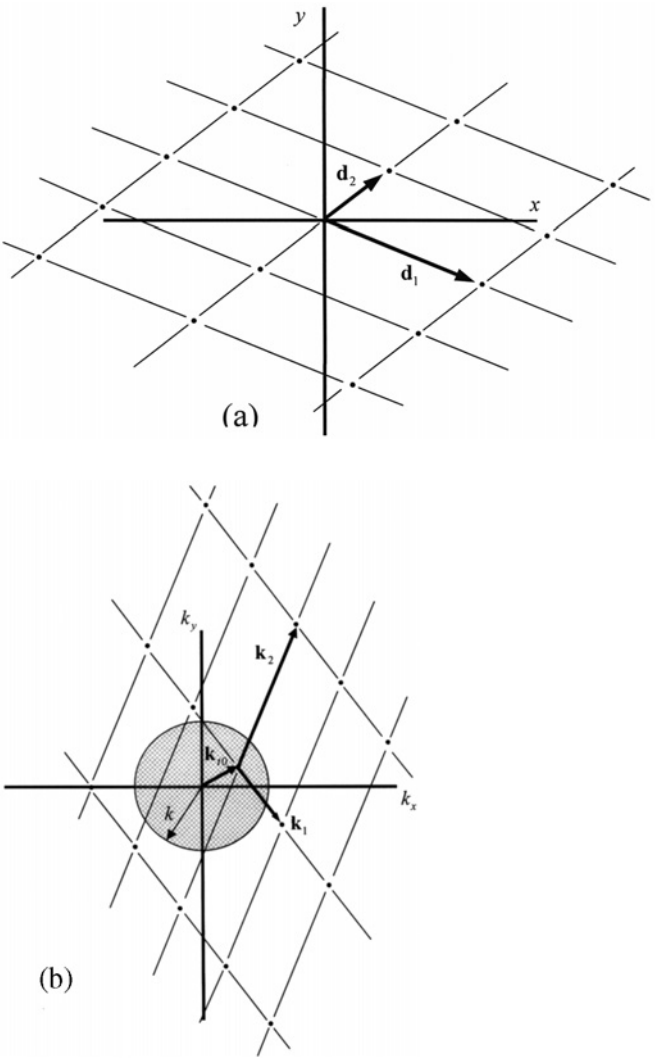


Figure 9.36 (a) The element lattice is defined by the lattice vectors $\mathbf{d}_1$ and $\mathbf{d}_2$. (b) The grating lobe lattice is shown translated by $\mathbf{k}_{t0}$, the phasing needed to scan the array beam within the circle representing the visible region. Lattice vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ are usually chosen such that no grating lobes appear within the visible region as $\mathbf{k}_{t0}$ varies over the desired scan range.

where

$$\mathbf{A}_{\text{ref}}(\mathbf{r}) = \frac{\mu_0}{4\pi r} e^{-jkr} \int_{S_{\text{ref}}} \mathbf{J}_{0,0}(\mathbf{r}') e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'} dS' \tag{9.104}$$

The far-zone electric field is proportional to the components of the vector potential tangent to the far-field sphere:

$$\mathbf{E}(\mathbf{r}) = -j\omega\left(A_\theta(\mathbf{r})\hat{\boldsymbol{\theta}} + A_\phi(\mathbf{r})\hat{\boldsymbol{\phi}}\right) \tag{9.105}$$

© 2006 by Taylor & Francis Group, LLC

The angle dependent factor for the reference element,

$$\begin{aligned}\mathbf{F}_{\text{ref}}(\hat{\mathbf{r}}) &= -\frac{4\pi r e^{jkR}}{j\omega\mu_0}\mathbf{E}_{\text{ref}}(\hat{\mathbf{r}}) \\ &= (\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}) \cdot \int_{S_{\text{ref}}} \mathbf{J}_{0,0}(\mathbf{r}') e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'} dS'\end{aligned} \tag{9.106}$$

is defined as the *element pattern* of the array. The element pattern incorporates, in principle, all the element's directional radiation, polarization, and mutual coupling characteristics. The array factor,

$$\text{AF}(\hat{\mathbf{r}}) = \sum_{m,n} a_{mn} e^{j\left[k\hat{\mathbf{r}}\cdot(m\mathbf{d}_1 + n\mathbf{d}_2) + \phi_{mn}\right]} \tag{9.107}$$

accounts for effects due to the array element configuration and excitation. The unit vector in the observation direction (θ, ϕ) is

$$\hat{\mathbf{r}} = \sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}} \tag{9.108}$$

To scan the array to a prescribed beam-pointing angle (θ_0, ϕ_0), the phase factor is chosen so that all contributions to the array factor add in phase in that direction,

$$\phi_{mn} = -k\hat{\mathbf{r}}_0 \cdot (m\mathbf{d}_1 + n\mathbf{d}_2) \tag{9.109}$$

where the unit vector in the beam direction is

$$\hat{\mathbf{r}}_0 = \sin\theta_0\cos\phi_0\hat{\mathbf{x}} + \sin\theta_0\sin\phi_0\hat{\mathbf{y}} + \cos\theta_0\hat{\mathbf{z}} \tag{9.110}$$

Defining observation and phasing wave vectors, $\mathbf{k} = k\hat{\mathbf{r}}$ and $\mathbf{k}_0 = k\hat{\mathbf{r}}_0$, respectively, the array factor may thus be succinctly written as

$$\text{AF}(\hat{\mathbf{r}}) = \sum_{m,n} a_{mn} e^{j(\mathbf{k}-\mathbf{k}_0)\cdot(m\mathbf{d}_1 + n\mathbf{d}_2)} \tag{9.111}$$

Components of the wavevectors in the plane of the array are more conveniently expressed in terms of the so-called *grating lobe lattice wave vectors*,

$$\mathbf{k}_1 = -\frac{2\pi(\hat{\mathbf{z}} \times \mathbf{d}_2)}{A} \qquad \mathbf{k}_2 = \frac{2\pi(\hat{\mathbf{z}} \times \mathbf{d}_1)}{A} \tag{9.112}$$

that are *biorthogonal* to the configuration space lattice vectors, i.e.,

$$\begin{aligned}&\mathbf{k}_1 \cdot \mathbf{d}_1 = 2\pi \qquad \mathbf{k}_1 \cdot \mathbf{d}_2 = 0, \\ &\mathbf{k}_2 \cdot \mathbf{d}_1 = 0 \qquad \mathbf{k}_2 \cdot \mathbf{d}_2 = 2\pi, \\ &A = \hat{\mathbf{z}} \cdot (\mathbf{d}_1 \times \mathbf{d}_2) = \text{element area}\end{aligned} \tag{9.113}$$

© 2006 by Taylor & Francis Group, LLC

From these properties, it is clear that if the components of the wave vectors transverse to the array normal (denoted by subscript "t") satisfy

$$\mathbf{k}_t = \mathbf{k}_{t0} + p\mathbf{k}_1 + q\mathbf{k}_2 \tag{9.114}$$

in Eq. (9.111) for integer values (p, q) and for any observable angles in the so-called *visible region*, $|k\hat{\mathbf{r}}| < k$, then in addition to the main beam, $(p = 0,\ q = 0)$, additional maxima called *grating lobes* of the array factor appear in the visible region, Fig. 9.36b. These undesired maxima thus appear whenever there exist nonvanishing integers (p, q) satisfying

$$\begin{aligned} &|\mathbf{k}_{t0} + p\mathbf{k}_1 + q\mathbf{k}_2| < k \quad \text{or} \\ &\left| k\left(\sin\theta_0 \cos\phi_0\, \hat{\mathbf{x}} + \sin\theta_0 \sin\phi_0\, \hat{\mathbf{y}}\right) + \frac{2\pi\hat{\mathbf{z}}}{A} \times (q\mathbf{d}_1 - p\mathbf{d}_2) \right| < k \end{aligned} \tag{9.115}$$

Equation (9.115) is used in phased array design to determine the allowable element spacings such that no grating lobes are visible when the array is scanned to the boundaries of its prescribed scan range in the angles (θ_0, ϕ_0) and at the highest frequency of array operation. The scan wave vector is shown in Fig. 9.37 and a graphical representation of these grating lobe conditions is depicted in Fig. 9.38. Scanning past one of the grating lobe circle boundaries overlapping the visible region allows a grating lobe into the visible region. For rectangular lattices, it is usually sufficient to check this condition only along the principal scan planes of the array. Triangular element lattices, however, often permit slightly larger spacings (hence larger element areas and fewer elements for a given gain requirement) than rectangular spacings [49]. For one-dimensional (linear) arrays, Eq. (9.115) reduces to

$$\frac{d}{\lambda} < \frac{1}{1 + |\sin\theta_0|} \tag{9.116}$$

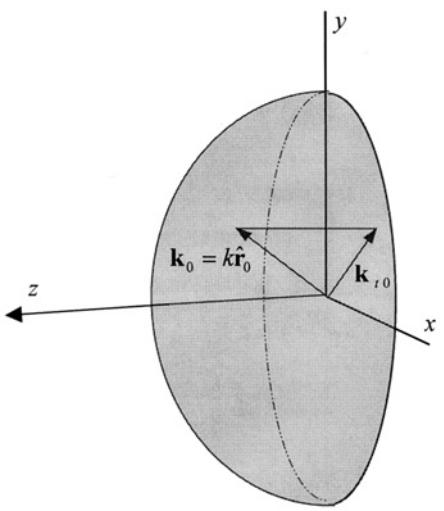


Figure 9.37 The vector $\mathbf{k}_0$ is in the desired scan direction; its projection, $\mathbf{k}_{t0}$, yields the transverse phase gradient required to scan the array to this direction.

© 2006 by Taylor & Francis Group, LLC

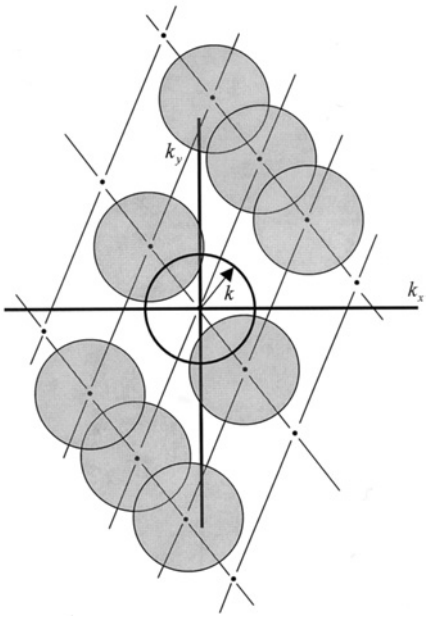


Figure 9.38 Grating-lobe lattice. The circle centers correspond to grating-lobe positions when the main beam is at broadside. As the array main beam is scanned within the visible region (the central unshaded circle), the grating lobes scan within the shaded circles. Grating-lobes enter the visible region when the main beam is scanned to a point in a shaded region overlapping the visible region circle.

where θ_0 is the maximum scan angle measured from the array axis normal. For rectangular element spacings, Eq. (9.116) may also be used to independently determine element spacings in the two orthogonal array planes. Note that for such arrays, grating lobes cannot appear if the spacings are less than $\lambda/2$ in each dimension; for 1λ spacings, grating lobes are at end fire and *always* appear in the visible region as the array is scanned. The above conditions merely ensure that the *peaks* of the grating lobe beams do not appear at end fire, i.e. in the plane of the array, when the array is scanned to a boundary of its scan coverage volume. To ensure that no part of a grating lobe beam appears in the visible region, slightly smaller element spacings than determined by the above procedure must be used. Not only the pattern but also the element impedance matching characteristics of an array may deteriorate rapidly at the onset of a grating lobe.

9.5.2. Array Pattern Characteristics

To examine array pattern characteristics more closely, we specialize to a rectangular array of $M \times N$ elements arranged in a rectangular lattice,

$$\mathbf{d}_1 = d_x\hat{\mathbf{x}} \qquad \mathbf{d}_2 = d_y\hat{\mathbf{y}} \qquad \mathbf{k}_1 = \frac{2\pi}{d_x}\hat{\mathbf{x}}, \qquad \mathbf{k}_2 = \frac{2\pi}{d_y}\hat{\mathbf{y}} \tag{9.117}$$

© 2006 by Taylor & Francis Group, LLC

The excitation of such an array is often *separable*, $a_{mn} = a_m b_n$, so that the array factor is also separable:

$$\mathrm{AF}(\hat{\mathbf{r}}) = \mathrm{AF}_x(\psi_x)\mathrm{AF}_y(\psi_y) \tag{9.118}$$

where

$$\psi_x = kd_x(\sin\theta\cos\phi - \sin\theta_0\cos\phi_0) \qquad \psi_y = kd_y(\sin\theta\sin\phi - \sin\theta_0\sin\phi_0) \tag{9.119}$$

and the array factor is the product of two linear array factors,

$$\mathrm{AF}_x(\psi_x) = \sum_n a_n e^{jn\psi_x} \qquad \mathrm{AF}_y(\psi_y) = \sum_m b_m e^{jm\psi_y} \tag{9.120}$$

If the array is uniformly excited, $a_m = 1/M, b_n = 1/N$, and centered with respect to the coordinate origin with M elements along the x dimension and N elements along the y dimension, the array factors can be summed in closed form, yielding

$$\mathrm{AF}_x(\psi_x) = \frac{\sin(M\psi_x/2)}{M\sin(\psi_x/2)} \qquad \mathrm{AF}_y(\psi_y) = \frac{\sin\left(N\psi_y/2\right)}{N\sin\left(\psi_y/2\right)} \tag{9.121}$$

Thus the array factor is a product of linear array factors of the form

$$\mathrm{AF}(\psi) = \frac{\sin(N\psi/2)}{N\sin(\psi/2)} \tag{9.122}$$

where

$$\psi = kd(\sin\alpha - \sin\alpha_0) \tag{9.123}$$

and α is the observation angle measured from a normal to the equivalent linear array axis; the linear array is assumed to be scanned to an angle α_0 from the array normal. The magnitude of $\mathrm{AF}(\psi)$ versus the parameter ψ for several values of N appears in Fig. 9.39. Since the array patterns are symmetric about $\psi = 0$ and periodic with period 2π, it suffices to plot them on the interval $(0, \pi)$. The pattern has a main beam at $\psi = 0$ with grating lobes at $\psi = \pm p2\pi = 1, 2, \ldots$. The scan angle and element spacing determine whether the grating lobes are visible. Between the main beam and first grating lobe, the pattern has $N - 1$ zeros at $\psi = p2\pi/N, p = 1, 2, \ldots, N - 1,$ and $N-2$ side lobes with peaks located approximately at $\psi = (p + 1/2)2\pi/N, p = 1, 2, \ldots, N - 2$. For large N, the first side-lobe level is independent of N and equal to that of a continuous, uniform aperture distribution, -13.2 dB. Indeed, in the vicinity of the main beam at $\psi \approx 0$, for large N we have

$$\mathrm{AF}(\psi) \approx \frac{\sin(N\psi/2)}{N(\psi/2)} \tag{9.124}$$

i.e., the array factor approximates the pattern of a uniform, continuous line source of length Nd. This observation holds for other array distributions as well: if the a_n are chosen

© 2006 by Taylor & Francis Group, LLC

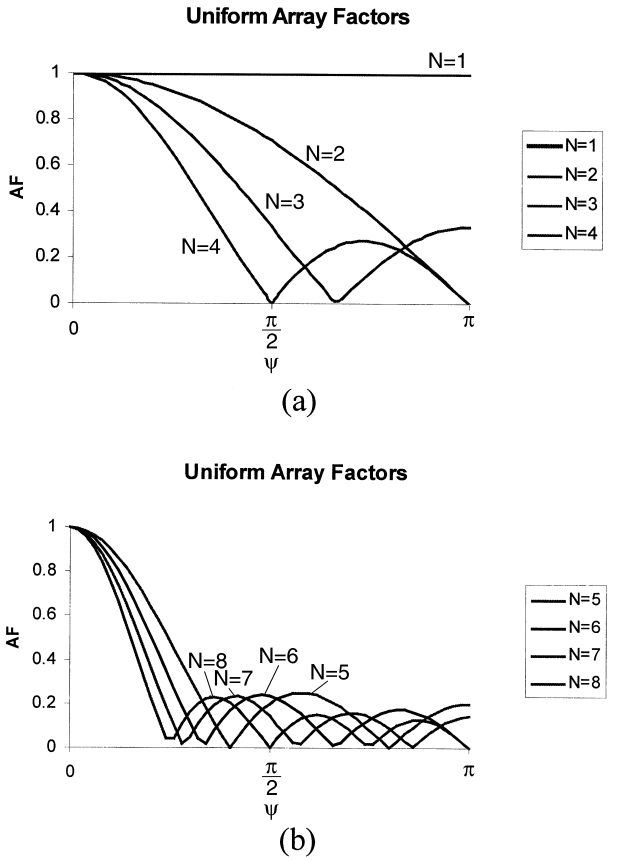


Figure 9.39 Array factors of uniform arrays with (a) $N = 1, 2, 3, 4$ and (b) $N = 5, 6, 7, 8$ elements.

as equi-spaced samples of a continuous distribution, then for large N the pattern in the vicinity of the main beam approaches that associated with the underlying sampled continuous distribution. Thus beam characteristics for continuous distributions may be used to estimate the corresponding quantities for array factors. Table 9.1 [50,51] lists approximate half-power beam widths, first side-lobe levels, and aperture efficiencies for large N for arrays whose element amplitude distributions are discrete samples of the listed continuous distributions. One of the more useful array distributions is the Taylor distribution [52], which allows one to choose $\bar{n}$ side lobes of specified and equal level on either side of the main beam, the remaining side lobes following the behavior of a uniform array pattern. Desirable features of the Taylor distribution are that it produces a physically realizable pattern that has essentially the narrowest beam possible for a given side-lobe level. Figures 9.40 and 9.41 show the pattern and corresponding element weights a_n, respectively, for a Taylor distribution with $\bar{n} = 6$ and a specified 20-dB side-lobe level. A convenient graphical method exists for determining a polar plot of the equivalent line source radiation pattern vs. the observation angle α for a given scan angle α_0. As illustrated for the uniform aperture distribution in Fig. 9.42, the pattern is first plotted vs. the parameter ψ and then projected onto a polar plot of diameter $2kd$ centered at $\psi = -kd \sin \alpha_0$. As the scan angle α_0 changes, the projected beam "scans," and the

© 2006 by Taylor & Francis Group, LLC

Table 9.1 Pattern Characteristics of Various Linear Array Aperture Distributions [50,51]

Aperture distribution $\lvert z_n\rvert < 1,\ n = 1, 2, \ldots, N$ $z_n = \dfrac{2(n-1) - N}{N}$	Aperture efficiency η_a	Half-power beam-width (degrees) L = array length	Maximum side-lobe level (dB below maximum)
Uniform: $a_n = 1$	1.00	51λ/L	13.2
Cosine: $a_n = \cos^n\left(\dfrac{\pi z_n}{2}\right)$			
$n = 0$	1	51λ/L	13.2
$n = 1$	0.81	69λ/L	23
$n = 2$	0.667	83λ/L	32
$n = 3$	0.575	95λ/L	40
Parabolic: $a_n = 1 - (1 - \Delta)z_n^2$			
$\Delta = 1.0$	1.0	51λ/L	13.2
$\Delta = 0.8$	0.994	53λ/L	15.8
$\Delta = 0.5$	0.970	56λ/L	17.1
$\Delta = 0$	0.833	66λ/L	20.6
Triangular: $a_n = 1 - \lvert z_n\rvert$	0.75	73λ/L	26.4
Cosine-squared on a pedestal: $a_n = 0.33 + 0.66\cos^2\left(\dfrac{\pi z_n}{2}\right)$	0.88	63λ/L	25.7
$a_n = 0.08 + 0.92\cos^2\left(\dfrac{\pi z_n}{2}\right)$	0.74	76.5λ/L	42.8

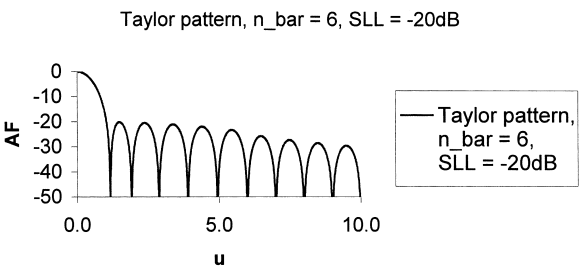


Figure 9.40 Pattern of a continuous Taylor aperture distribution: $\bar{n} = 6$, SLL $= -20$ dB.

dependence of grating lobe onset on scan angle and element spacing becomes apparent. Further, the approximate broadening of the main beam with scan angle by a factor $1/\cos\alpha_0$ also becomes apparent. For a line source of omnidirectional elements, the actual pattern is obtained by rotating the pattern of the figure about the horizontal axis to form a conical beam.

For separable rectangular aperture distributions, such a graphical representation is not practical to construct, yet it is convenient for visualization. We first imagine that

© 2006 by Taylor & Francis Group, LLC

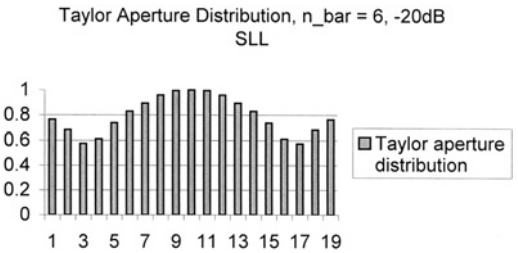


Figure 9.41 Sampled values of continuous Taylor aperture distribution for a 19-element array: $\bar{n} = 6$, SLL $= -20$ dB.

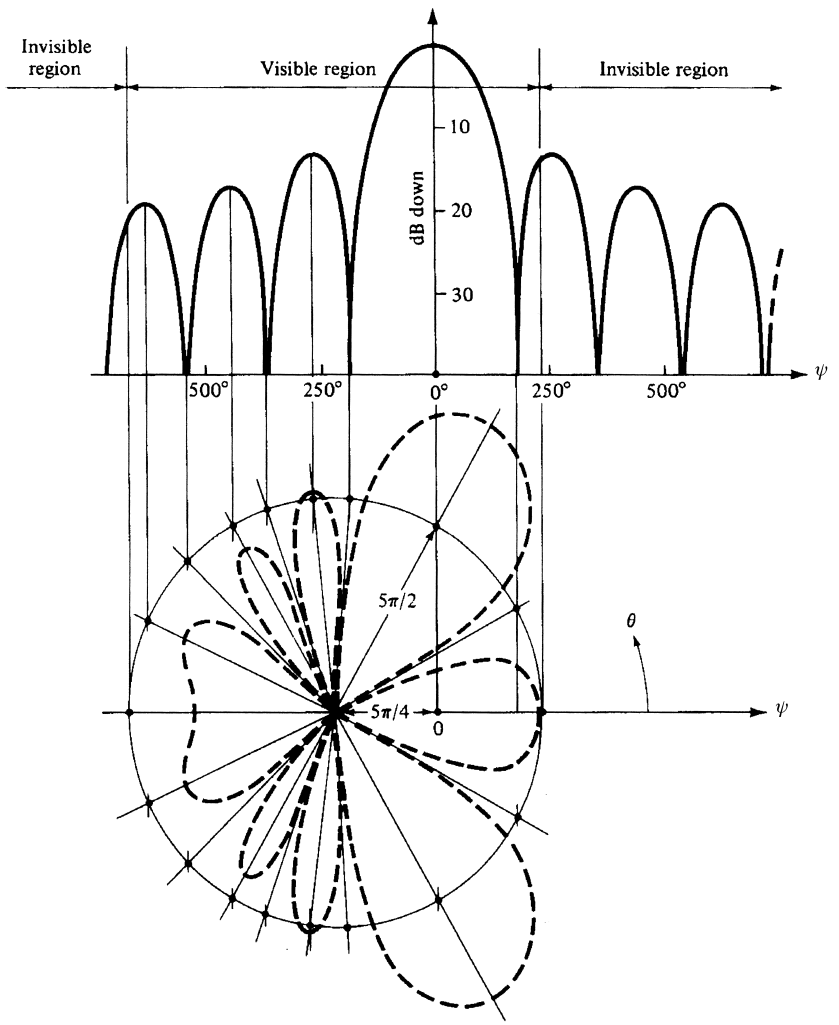


Figure 9.42 Projection of linear array pattern to obtain polar pattern. The location of the center of the polar pattern is determined by the interelement phase shift. The radius of the polar pattern, and hence the visible region, is determined by the frequency. The angle θ in the figure is $\pi/2 - \alpha$. (This material is used by permission of John Wiley and Sons, Inc., *Antenna Theory: Analysis and Design*, C. Balanis, 1997.)

© 2006 by Taylor & Francis Group, LLC

the product $AF_x(\psi_x)AF_y(\psi_y)$ is plotted vs. ψ_x/d_x and ψ_y/d_y, where, because of the separability property, pattern cuts for any constant ψ_x/d_x or ψ_y/d_y are identical within a scaling factor. A hemisphere of radius k is then centered at $\psi_x/d_x = -k\sin\theta_0\cos\phi_0$, $\psi_y/d_y = -k\sin\theta_0\sin\phi_0$, and the array-factor pattern multiplied by the vector-valued element pattern is projected onto the hemisphere to obtain the three-dimensional radiation pattern.

9.5.3. Array Gain

The array directivity is usually defined in terms of the *scan element pattern*. Since the directivity of an array changes with its scan and with the mismatch to its feed line, it is convenient to incorporate these effects into the definition of an element gain by referring it to the power *available* rather than the power *input* to the array. Consequently the scan element pattern (gain) is defined as

$$g_{\text{scan}}(\hat{\mathbf{r}}_0) = \frac{4\pi r^2 \left|\mathbf{E}_{\text{scan}}(\hat{\mathbf{r}}_0)\right|^2}{2\eta_0 P_{\text{avail}}} \tag{9.125}$$

where $\mathbf{E}_{\text{scan}}(\hat{\mathbf{r}}_0)$ is the far-zone electric field radiated by a single element in the direction $\hat{\mathbf{r}}_0$ with all other elements terminated in their generator impedance and P_{avail} is the power available to the element. Note that the quantity in Eq. (9.125) is defined in such a way that it is readily measurable. Since the radiated field is proportional to the excitation coefficient a_{mn}, and the total power available at each element is proportional to its square, the array gain is given by

$$G_{\text{array}}(\hat{\mathbf{r}}_0) = \frac{4\pi r^2 \left|\mathbf{E}_{\text{scan}}(\hat{\mathbf{r}}_0)\right|^2}{2\eta_0 P_{\text{avail}}} \frac{\left|\sum_{m,n} a_{mn}\right|^2}{\sum_{m,n} |a_{mn}|^2} = N_{\text{tot}}\, g_{\text{scan}}(\hat{\mathbf{r}}_0)\eta_a \tag{9.126}$$

where

$$\eta_a = \frac{\left|\sum_{m,n} a_{mn}\right|^2}{N_{\text{tot}} \sum_{m,n} |a_{mn}|^2}$$

$$\left[= \frac{\left|\sum_m a_m\right|^2 \left|\sum_n a_n\right|^2}{\left(M\sum_m |a_m|^2\right)\left(N\sum_n |a_n|^2\right)} = \eta_{xa}\eta_{ya} \qquad \text{for separable aperture distributions} \right] \tag{9.127}$$

is the *aperture efficiency*, and N_{tot} is the total number of array elements. The equivalent linear array aperture efficiencies, η_{xa}, η_{ya}, that can be used for separable planar apertures are tabulated for a number of common aperture distributions in Table 9.1.

© 2006 by Taylor & Francis Group, LLC

In the following we assume, without loss of generality, that the array excitation distribution is uniform, $a_{mn} = 1$, and relate two different forms of excitation. We define the *isolated element pattern*, $\mathbf{E}_{\text{iso}}(\hat{\mathbf{r}}_0)$, as the far-field element pattern resulting when the terminals of the element are driven by a current source I_{iso}, and all other elements are open-circuited at their terminals. For many elements such as dipoles, the open-circuited elements support negligible currents and hence the isolated element pattern is essentially that of a single element with the remaining array elements removed: $\mathbf{E}_{\text{iso}}(\hat{\mathbf{r}}_0) \approx \mathbf{E}_{\text{single}}(\hat{\mathbf{r}}_0)$. Since the array radiation may be expressed as an appropriate superposition over *either* the scan element pattern or the isolated element pattern, we have

$$N_{\text{tot}}\mathbf{E}_{\text{scan}}(\hat{\mathbf{r}}_0) = N_{\text{tot}}\frac{I_{\text{scan}}(\hat{\mathbf{r}}_0)\mathbf{E}_{\text{iso}}(\hat{\mathbf{r}}_0)}{I_{\text{iso}}} \tag{9.128}$$

where $I_{\text{scan}}(\hat{\mathbf{r}}_0)/I_{\text{iso}}$ is the ratio of current at the terminals of each element under scan conditions with all elements excited to the terminal current of the isolated element.

The power available per element is given by

$$P_{\text{avail}} = \frac{|V_g|^2}{8R_g} = \frac{|I_{\text{scan}}(\hat{\mathbf{r}}_0)|^2|Z_{\text{scan}}(\hat{\mathbf{r}}_0) + Z_g|^2}{8R_g} \tag{9.129}$$

where V_g is the generator voltage, $Z_g = R_g + jX_g$ is its internal impedance, and

$$Z_{\text{scan}}(\hat{\mathbf{r}}_0) = Z_{\text{iso}} + \sum_{\substack{m,n;\\ m=n\neq 0}} Z_{mn}e^{-jk\hat{\mathbf{r}}_0\cdot(m\mathbf{d}_1+n\mathbf{d}_2)} \tag{9.130}$$

For a large array, the scan impedance $Z_{\text{scan}}(\hat{\mathbf{r}}_0)$ is essentially that of an infinite array. In Eq. (9.130), $Z_{\text{iso}} = R_{\text{iso}} + jX_{\text{iso}}$ is the isolated element input impedance and Z_{mn} is the mutual impedance between elements separated by $m\mathbf{d}_1 + n\mathbf{d}_2$ in the array lattice.

The isolated element gain, in terms of quantities previously defined, is

$$g_{\text{iso}}(\hat{\mathbf{r}}_0) = \frac{4\pi\left(r^2|\mathbf{E}_{\text{iso}}(\hat{\mathbf{r}}_0)|^2/2\eta_0\right)}{(1/2)|I_{\text{iso}}|^2R_{\text{iso}}} \tag{9.131}$$

Combining Eqs. (9.125), (9.128), (9.129), and (9.131), we have

$$g_{\text{scan}}(\hat{\mathbf{r}}_0) = \frac{4R_gR_{\text{iso}}g_{\text{iso}}(\hat{\mathbf{r}}_0)}{|Z_{\text{scan}}(\hat{\mathbf{r}}_0) + Z_g|^2} \tag{9.132}$$

Defining a conjugate reflection coefficient,

$$\Gamma_*(\hat{\mathbf{r}}_0) = \frac{Z^*_{\text{scan}}(\hat{\mathbf{r}}_0) - Z_g}{Z_{\text{scan}}(\hat{\mathbf{r}}_0) + Z_g} \tag{9.133}$$

© 2006 by Taylor & Francis Group, LLC

Eq. (9.132) can be written as

$$g_{\text{scan}}(\hat{\mathbf{r}}_0) = \frac{R_{\text{iso}}}{R_{\text{scan}}(\hat{\mathbf{r}}_0)} g_{\text{iso}}(\hat{\mathbf{r}}_0)\left(1 - \left|\Gamma_*(\hat{\mathbf{r}}_0)\right|^2\right) \tag{9.134}$$

This result shows how the scan element pattern and the isolated element pattern are linked through the element mismatch and scan element resistance. If the array is assumed matched at angle $\hat{\mathbf{r}}_{\text{match}}$, i.e., $\Gamma_*(\hat{\mathbf{r}}_{\text{match}}) = 0, \quad R_{\text{scan}}(\hat{\mathbf{r}}_{\text{match}}) = R_g$, then Eq. (9.134) may be expressed as

$$g_{\text{scan}}(\hat{\mathbf{r}}_0) = \frac{R_{\text{g}} g_{\text{scan}}(\hat{\mathbf{r}}_{\text{match}})}{R_{\text{scan}}(\hat{\mathbf{r}}_0)} \frac{g_{\text{iso}}(\hat{\mathbf{r}}_0)}{g_{\text{iso}}(\hat{\mathbf{r}}_{\text{match}})} \left(1 - \left|\Gamma_*(\hat{\mathbf{r}}_0)\right|^2\right) \tag{9.135}$$

Equation (9.134) is particularly convenient for use when the isolated element pattern is almost identical to the single element pattern, i.e., when the currents on the unexcited element essentially vanish when their terminals are *opened*. But this is not the case for some elements, such as slots and patch antennas. For such elements, *shorting* the unexcited elements terminals in an isolated element pattern renders the pattern essentially the same as the single element pattern. Repeating the derivation under the assumption that the isolated pattern is that of a singly excited element with all others terminated in *short* circuits leads to

$$g_{\text{scan}}(\hat{\mathbf{r}}_0) = \frac{G_{\text{iso}}}{G_{\text{scan}}(\hat{\mathbf{r}}_0)} g_{\text{iso}}(\hat{\mathbf{r}}_0)\left(1 - \left|\tilde{\Gamma}_*(\hat{\mathbf{r}}_0)\right|^2\right) \tag{9.136}$$

where G_{iso} and $G_{\text{scan}}(\hat{\mathbf{r}}_0)$ are the isolated and scan conductances, respectively, and a new conjugate reflection coefficient is defined as

$$\tilde{\Gamma}_*(\hat{\mathbf{r}}_0) = \frac{Y_g - Y^*_{\text{scan}}(\hat{\mathbf{r}}_0)}{Y_g + Y_{\text{scan}}(\hat{\mathbf{r}}_0)} \tag{9.137}$$

Hansen [53] points out that the approximate result frequently appearing in the literature,

$$g_{\text{scan}}(\hat{\mathbf{r}}_0) = \frac{4\pi A_{\text{ref}} \cos\theta_0}{\lambda^2}\left(1 - \left|\Gamma(\hat{\mathbf{r}}_0)\right|^2\right) \tag{9.138}$$

where A_{ref} is the area of a unit cell and $\Gamma(\hat{\mathbf{r}}_0)$ is the terminal reflection coefficient, holds only when no grating lobes or higher-order feed modes are present and the elements are thin and straight.

In many arrays, dielectric slabs or substrates on ground planes are used, and these may excite a surface wave when the transverse scan wave number equals the surface wave propagation constant. The energy in the surface wave is not available for radiation, and hence the effect on the directivity is seen as a "blind spot," i.e., a large reflection coefficient or sharp dip in the directivity pattern. Similarly, the onset of a grating lobe may also represent a reduction in element directivity and hence a large reflection coefficient.

© 2006 by Taylor & Francis Group, LLC

9.5.4. Array Elements

Almost any electrically small radiator may be used in a phased array, but because of their low cost and/or ease of fabrication, typical elements include dipoles, slots, open-ended waveguides, patches, and notch antennas [53]. Because of strong mutual coupling effects in the array environment, it cannot be expected that the behavior of these elements resembles that of the corresponding isolated element. To account for these coupling effects, calculations of element active impedances are often performed for elements in *infinite* array environments. Unfortunately, array elements spaced about one-half wavelength apart often do not act as if in an infinite array environment unless they are located about 10 elements away from the array edges. Thus in a 60×60 array only about 44% of the elements satisfy this condition. Nevertheless, the infinite array active impedance is frequently used to predict element behavior in the phased array environment.

Wire or *printed-circuit dipoles* are often used at low frequencies. The dipole arms are usually about a half wavelength in length and located approximately a quarter wavelength above a ground plane to direct radiation into the forward direction. Bending the arms of the dipole towards the ground plane can increase the angular coverage; using thicker elements tends to increase both the bandwidth and reduce mutual coupling. Printed-circuit dipoles [54] are popular because of the ease with which they may be fabricated; two such dipoles placed orthogonal to one another and fed with a 90° phase difference can provide circular polarization.

Slots milled in the sides of waveguide walls are often used in applications where high power or accurate control of manufacturing for low side-lobe level designs is required. Unless the slots couple into another waveguide or transmission line containing phase shifters, the interelement phase is not generally controllable along the guide dimension. However, if the slots are cut into the narrow wall of the guide, a series of slotted waveguides may be stacked closely together to avoid grating lobes, and phasing between the stacked guides may be introduced to effect scanning in the orthogonal plane. Narrow wall slots cut perpendicular to the waveguide edges do not couple energy from the guide, and hence the slots must be tilted, the angle of tilt determining the degree of coupling. Alternating the tilt angle of adjacent slots alternates the sign of the coupling so that in-phase excitation results when the slots are placed $\lambda_g/2$ apart, where λ_g is the guide wavelength. The slot coupling should be small so as to minimize reflections; therefore a certain fraction of the energy at the input to the guide remains after the last slot. This remaining energy must be absorbed in a matched load, thus reducing the antenna efficiency. For arrays with tapered aperture distributions and many slots, where it is relatively easy to use slots with small coupling, it may be possible to reduce this loss to as low as 1% or 2%. In any case, slot array design is often a tradeoff between dealing with slot maximum coupling limitations and expending energy in the matched load at the end of the line.

Open-ended waveguides are often used as radiators because the waveguides can easily accommodate bulky phase shifters. The resultant structures are not only relatively easy to analyze and match but are also mechanically strong and capable of handling high power. Furthermore, they are suitable for applications requiring flush mounting. The waveguides may also be loaded with dielectrics to reduce the element size. This may be needed both to avoid grating lobes and to provide sufficient space to assemble the structure. Dielectric slabs are often used to match the guides to free space, but they may also support the propagation of surface waves along the array face.

Patch antennas consisting of a thin metallic layer bonded to a grounded substrate are popular array elements because of their ease of fabrication, light weight, low profile, and

© 2006 by Taylor & Francis Group, LLC

ability to conform to a planar or curved surface [54–56]. They may be either probe fed by extending the center conductor of a coaxial line through the ground plane and attaching it to the patch, by a microstrip line coupled either directly or indirectly (proximity coupled) to the patch, or aperture coupled to the feeding microstrip line below the aperture in the ground plane. Microstrip patch elements may have strong mutual coupling and narrow bandwidth.

Flared notch antennas may be thought of as slots in a ground plane that are flared to form a one-dimensional horn shape. The gradual curve allows for a broadband match for these elements [57].

9.5.5. Phased Array Feed and Beam-forming Systems

A major concern in phased array design is distributing radiated energy to or collecting received energy from the array elements. Since feed networks for transmission are generally reciprocal, they can, of course, also serve as beam-forming networks for receiving functions. But received energy from phased arrays is frequently formed into multiple, simultaneous beams and this capability is not generally needed for transmitting. For this reason, and because it is often desirable to isolate transmission and receive functions, beam-forming networks are often separated from the transmission feed network. In modern phased arrays, multiple beams are typically generated by digital rather than analog processing, and hence we concentrate on feed networks for generating single transmitting beams. Most phased arrays can be classified as parallel, series, space, or active aperture fed systems. Hybrid systems may employ one type of feed system along one dimension of a planar array and another along the other dimension.

Parallel feed systems are often called *corporate feed* systems because of their resemblance to a corporate organization chart. As illustrated in Fig. 9.43, it is desirable to employ feed networks with a branching network of hybrid junctions that absorb reflections from the elements. These prevent reflections from being reradiated by the array and resulting in pattern deterioration. Corporate feeds for microstrip patch arrays often employ only power dividers, however. Corporate feed systems are generally simpler to design than series feed systems since each element excitation can be controlled independently and all transmitter-to-element paths are equal in length and hence involve the same phase differences and path losses.

Series feed systems couple energy at periodically spaced locations along a guiding system such as a waveguide, as shown in Fig. 9.44. The electrical length between tap positions is frequency dependent and causes series fed arrays to naturally scan slightly with frequency. This tendency is purposely enhanced in frequency scanning arrays by folding the feed line between tap points to further increase the electrical length—at the expense of further lowering the system's bandwidth. Directional couplers may be used to isolate reflections that occur at the many tap locations, and element locations may be chosen to differ slightly from λ_g or $\lambda_g/2$ spacings so that reflections do not add in phase, resulting in

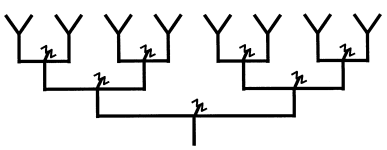


Figure 9.43 Corporate feed for an eight-element array with hybrids at each feed-line junction.

© 2006 by Taylor & Francis Group, LLC

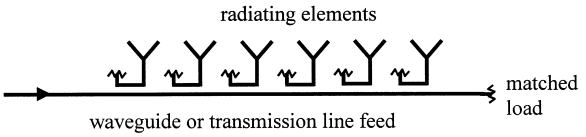


Figure 9.44 Series feed array using directional couplers.

a large input VSWR. The design of series feed systems is iterative [58], as may be seen by considering a common configuration: An array of slots milled into the narrow wall of a waveguide. The slot angle relative to the guide edges determines the degree of coupling for each slot, and there is limit to the maximum practical coupling, especially since the slot conductances should be small to minimize reflections. The aperture distribution a_n, number of slots N, and maximum allowed conductance eventually determine $r = P_{\text{load}}/P_{\text{in}}$, the fraction of the input power remaining in the guide after the last slot, which must be dissipated in a matched load at the end of the feed line. The iterative design process begins with an assumed value for r. Neglecting reflections, the fractional power dissipated by slot n is proportional to $F_n = a_n^2$. Beginning at the load end of the feed line, the design proceeds to determine all the slot conductances for the assumed r. For slots with low reflections on a lossless line, the normalized conductances can be approximately determined from [58]

$$g_n = \frac{F_n}{[1/(1-r)]\sum_{m=1}^{N} F_m - \sum_{m=1}^{n} F_m} \tag{9.139}$$

where element N is nearest the load. If a slot conductance is generated that exceeds the allowed maximum, then r must be increased and the procedure repeated; if no conductance values near the maximum allowable are found, then r may be reduced to increase the array efficiency. The final slot conductances then determine the slot angles. The approach may also be generalized to account for a line attenuation factor between elements [53]. In a series fed array, it is not unusual to dissipate 5–10% of the array input energy in the terminating load.

Lens arrays are fed by illuminating the elements of a receiving aperture by an optical feed system and then retransmitting the received signal after passing it through phase shifters that serve not only to scan the beam but also to correct for the differing path lengths taken from the optical feed to the transmitting aperture. The primary feed is often a single horn antenna or a cluster of horns forming a monopulse system. The feed system's name reflects the fact that the phase shifters act as a lens to collimate the transmitted beam—or since the system is reciprocal, to focus it on receive. To reduce the adverse effect that reflections from the lens have on the VSWR of the optical feed, the feed is often offset. An advantage of the approach is its relative simplicity and low cost. A disadvantage is the fact that antenna elements are required on both faces of the array.

Reflect arrays are similar to lens arrays, except that there is only one array face containing antenna elements. Energy from the optical feed is collected by the elements at the aperture, passes through phase shifters, is reflected by a short circuit, passes back through the phase shifters and is reradiated as a scanned beam plane at the aperture plane. Since the signal passes through the phase shifters twice, the phase shift settings are only

© 2006 by Taylor & Francis Group, LLC

half the total needed and the phase shifters must be of the reciprocal type. Lens and reflect arrays share not only the advantages, but also the disadvantages of other optically fed systems. One has less control over the array aperture distribution, which is controlled primarily by feed pattern. Spillover is also a concern for both feed types, and reflect arrays therefore generally employ an offset feed to reduce feed blockage.

None of the feed systems described have the power handling capabilities of *active-aperture* systems in which a transmit–receive module is associated with—and possibly contains—each array element. The element module combines transmit–receive switches, solid-state transmitter and receiver amplifiers, and phase shifters. Feed losses are thus practically eliminated, and, since phase shifters may be located at the transmitter front end, they are not required to handle high power. On receive, not only is the signal-to-noise ratio unaffected by feed or other losses, but also digital beam combining of the receiver outputs may be used to control receive patterns, including adaptively controlling sidelobe levels and null positions.

Monolithic integrated phased arrays attempt to lower the cost per element and increase the reliability and repeatability of a phased array by combining many elements, their transmitter and receive functions, as well as beamforming and array control functions, on a single package. So-called *brick* configurations combine elements common to a row and use the depth dimension to accommodate array components and feed structures. *Tile* configurations combine a number of elements in the same plane with various array components located in separate, parallel layers.

9.5.6. Electronic Beamsteering

The phase of a signal traveling through a wave guiding section of length ℓ with cutoff frequency f_c (in a TEM line, $f_c = 0$) is

$$\varphi = 2\pi\ell\sqrt{\mu\varepsilon\left(f^2 - f_c^2\right)} \tag{9.140}$$

To change this phase so as to scan an array beam, one may change

1. The line length by switching between different line lengths (diode phase shifters)
2. The permeability μ (ferrite phase shifters), or less frequently, the permittivity ε (ferroelectric and plasma phase shifters)
3. The frequency f (frequency scanning)
4. The guide cutoff frequency.

All these approaches have been used, including changing the guide cutoff frequency by mechanically changing the guide dimension, though the latter cannot be considered *electronic beam steering*.

At frequencies below *S* band, diode phase shifters generally have less loss than ferrite phase shifters; above *S* band, ferrite phase shifters are usually preferred. Phase shifters are usually digitally controlled and provide quantized values of phase shift, the resolution being determined by the smallest bit of the phase shifter. An N bit phase shifter provides phase shifts between $0^\circ$and 360° of phase in steps of $360^\circ/2^N$. Since the desired phase variation is linear whereas achievable phase settings are discrete, the phase error due to quantization is a periodic ramp function. This periodic phase error produces *quantization*

© 2006 by Taylor & Francis Group, LLC

grating lobes, the *n*th one of whose magnitude, normalized to the peak of the array factor, is given by

$$\frac{1}{|n|2^N} \qquad n = \pm 1, \pm 2, \ldots \tag{9.141}$$

for an N bit phase shifter. The angular locations of these quantization lobes depend on the scan angle and number of phase bits, and most will not appear in the visible region of the pattern. Quantization errors also result in beam pointing errors and reduced gain. The "round-off error" in phase due to quantization may be randomized by introducing a known but random phase offset for each element; the resulting element phase errors are incoherent and therefore spread out the grating lobe energy as a random background quantization noise added to the designed array pattern. Thus quantization grating lobes are eliminated at the expense of raising the RMS sidelobe level [59].

Diode phase shifters generally operate as switches that change the electrical length of a signal path according to whether the switch, usually a PIN diode, is in the open or closed state. The two most commonly used types are switched and hybrid-coupled phase shifters. As illustrated in Fig. 9.45, the switched type uses diodes to switch different line lengths into the feed line to obtain the required phase shift. In a hybrid-coupled phase shifter, one bit of which is shown in Fig. 9.46, the states of the two diodes in the branch lines of the hybrid are the same and determine varying reflection point locations along the lines.

Ferrite phase shifters of the latching type operate by changing the magnetization state, and hence the insertion phase, of a ferrite toroid. These phase shifters are commonly employed in waveguide feeds, and usually consist of several cascaded sections of differing lengths—each length representing a different phase shifter bit—to obtain the desired phasing. As shown in Fig. 9.47, the magnetization is controlled by a current pulse provided by a wire threading each toroidal bit. Dielectric spacers between bits provide matching. The pulse drives the core into saturation and the remanence flux provides the magnetization required without need for a holding current. The resulting permeability change provides the phase shift. Such phase shifters are nonreciprocal and hence must be switched between the transmit and receive modes; for this reason, they cannot be used in reflect arrays since they cannot be switched quickly enough during the short time between the passage of incident and reflected pulses, whether operating in transmit or receive modes.

Frequency scanning systems do not require phase shifters but instead rely on changing the frequency to affect the electrical length, and hence the phase, between

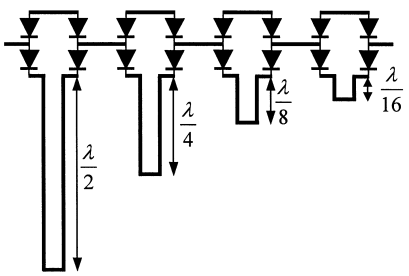


Figure 9.45 A four bit, digitally switched diode phase shifter. An N-bit phase shifter provides a phase increment of $360^\circ/2^N$ and requires $4N$ PIN diodes.

© 2006 by Taylor & Francis Group, LLC

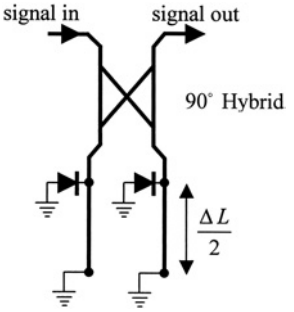


Figure 9.46 A phase bit using a hybrid coupler. Changing the diode states from conducting to nonconducting changes the round-trip path length to the reflection point by ΔL. An N bit phase shifter provides a phase increment of $360^\circ/2^N$ and requires $2N$ PIN diodes.

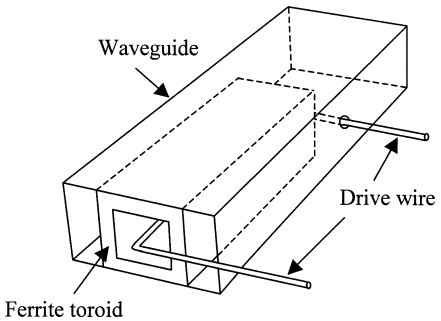


Figure 9.47 A single bit of a ferrite phase shifter in a waveguide. A current pulse in the drive wire saturates the ferrite core; reversing the direction of the pulse reverses the magnetization, thereby changing the phase shift of a signal traversing the core. Different phase-shift bit values are produced by cascading toroids of varying lengths.

elements excited by a traveling wave series feed [60]. At the expense of array bandwidth, the scan sensitivity to frequency (i.e., the change in scan angle per unit change in frequency) is enhanced by folding the feed line to increase the electrical length between elements. The resulting feeds are often called *sinuous* or *serpentine* feeds. If the main beam points at broadside at a frequency f_0, then at a frequency f the scan angle θ_0 is given by

$$\sin\theta_0 = \frac{L}{df}\left(\sqrt{f^2 - f_c^2} - \sqrt{f_0^2 - f_c^2}\right) \tag{9.142}$$

where L is the element separation distance as measured along a mean path inside a waveguide feed with guide cutoff frequency f_c and d is the actual element separation. Equation (9.142) also applies to TEM line feeds with $f_c = 0$, but, with less scan sensitivity, they are used less often. The factor L/d, which is often called the *wraparound* or *wrap-up factor*, controls the sensitivity. At broadside, reflections from element mismatches in the feed, though small, add in phase and may result in a large VSWR at the feed line input. For this reason, some frequency scan systems cover only an angular sector from a few degrees off broadside to nearly end fire. Because of their inherently low bandwidth

© 2006 by Taylor & Francis Group, LLC

characteristics, frequency-scanning systems are not often used in modern phased array systems.

9.5.7. Mutual Coupling

Not only is mutual coupling always present in phased arrays, but it is also responsible for most of their unique characteristics. Though it should not be neglected in array design, it is the parameter that is most difficult to obtain. Several general principles concerning mutual coupling are obvious, however [53]:

1. Coupling decreases with distance between elements.
2. Coupling between elements is strongest when their maximum radiation directions are aligned along their line of separation. For linear dipole elements in free space, this occurs, e.g., when they are parallel rather than collinear.
3. Coupling is smaller between large, highly directional elements such as horns.
4. Coupling is stronger between elements when their substrates support surface waves.

Many qualitative array effects can be discerned from the analysis of Wheeler [61] using an infinite sheet current model. The model has also been extended to include dielectric substrates [62] and demonstrates many of the most important effects of element spacing, polarization, scan angle, and substrates. A related concept is the grating-lobe series [63]. These and even more realistic models approach the problem from an infinite-array point of view. If an array is large and the taper is gradual, the interaction between central elements of the array may be approximated by those of an infinite array of like elements; the analysis is then reduced to that of a unit or reference cell of the array. Usually one determines the *scan impedance* or *scan reflection coefficient* of the reference element in the presence of an infinite number of elements similarly excited but with a fixed interelement phase shift. The calculation often requires numerical methods. In principle, integration of the resulting quantity over the phase shift variables on a unit cell of the grating lobe lattice yields the interaction between elements for a singly-excited element. This result forms the basic connection between infinite array analysis and analysis from the opposite extreme—element-by-element analysis.

Element-by-element analyses are necessary for small-to-moderate size arrays and benefit from knowledge of the interaction between isolated pairs of elements. For example, a convenient approximate form for the mutual impedance between linear dipoles in echelon is available [53,64] and is easily extended to slots in a ground plane. Limited data for mutual coupling between horns, open-ended waveguides, microstrip patches, and other array elements are also available or may be computed numerically.

Blind angles are angles at which the scan reflection coefficient is near unity or, equivalently, the scan element pattern is near zero. They may be interpreted, respectively, as angles for which higher modes cancel with the dominant mode in the element, or as angles that allow coupling to a leaky mode on the array [65]. The leaky mode is essentially a surface wave that is supported by the periodic array structure, which is leaky due to radiation from one of the space harmonics of the mode (the phase constant of the radiating space harmonic corresponds to the wave number of the array phasing at the blindness angle). If the periodic loading effect is not very strong, the phase constant of the leaky wave may be approximated as that of the corresponding surface wave. Surface waves have wave numbers $\beta_{sw} > k$, and if circles of radius β_{sw} are added to and centered on

© 2006 by Taylor & Francis Group, LLC

the grating lobe lattice, their intersections with the visible region circle locate possible angles where blind spots can occur [66,67].

Controlling mutual coupling is of utmost importance in phased array design, and several attempts have been made to either reduce or compensate for coupling effects. Grating lobe series analysis shows that close element spacing reduces the variation of reactance since grating lobes, which primarily affect reactance, are pushed further into the invisible region. H-plane baffles placed between rows of slots or dipoles have also been used to significantly reduce impedance variations in the two principal scan planes [68]. Approaches used on open-end waveguide arrays have included the control of multimode excitation in the unit cell by dielectric loading [69] and by adding irises [70]. Other alternatives include the use of slot arrays with parasitic monopoles [71] and of dielectric sheets. Several such sheets in cascade have been used to form a wave filter placed sufficiently far in front of the array so as not to affect impedance, but to improve the scan element pattern [72].

REFERENCES

1. Garg, R.; Bhartia, P.; Bahl, I.; Ittipiboon, A. *Microstrip Antenna Design Handbook*; Artech House: Norwood, MA, 2000.
2. Lee, K.F. Ed. *Advances in Microstrip and Printed Antennas*; Wiley: New York, 1997.
3. Pozar, D.M.; Schaubert, D.H. *Microstrip Antennas: The Analysis and Design of Microstrip Antennas and Arrays*; IEEE Press, Piscataway, NJ, 1995.
4. Gardiol, F.E. *Broadband Patch Antennas*; Artech House: Norwood, MA, 1995.
5. Jackson, D.R.; Williams, J.T. A comparison of CAD models for radiation from rectangular microstrip patches. Int. J Microwave Millimeter-Wave CAD **April 1991**, *1* (2), 236–248.
6. Jackson, D.R.; Long, S.A.; Williams, J.T.; Davis, V.B. Computer-aided design of rectangular microstrip antennas. In: *Advances in Microstrip and Printed Antennas*; Lee, K.F. Ed.; Wiley, New York, 1997; Chap. 5.
7. Pozar, D.M. A reciprocity method of analysis for printed slot and slot-coupled microstrip antennas, IEEE Trans. Antennas Propagation **Dec. 1986**, *AP-34*, 1439–1446.
8. Kumar, G.; Ray, K.P. Broadband Microstrip Antennas; Artech House, Norwood, MA, 2003.
9. Pues, H.; Van de Capelle, A. Accurate transmission-line model for the rectangular microstrip antenna, Proc. IEEE, Vol. 131, Pt. H, No. 6, pp. 334–340, Dec. 1984.
10. Richards, W.F.; Lo, Y.T.; Harrison, D.D. An improved theory of microstrip antennas with applications, IEEE Trans. Antennas Propagation **Jan. 1981**, *AP-29*, 38–46.
11. Pozar, D.M. Input impedance and mutual coupling of rectangular microstrip antennas. IEEE Trans. Antennas Propagation **Nov. 1982**, *AP-30*, 1191–1196.
12. Prior, C.J.; Hall, P.S. Microstrip disk antenna with short-circuited annular ring. Electronics Lett. **1985**, *21*, 719–721.
13. Guo, Y.-X.; Mak, C.-L.; Luk, K.-M.; Lee, K.-F. Analysis and design of L-probe proximity fed patch antennas, IEEE Trans. Antennas Propagation **Feb. 2001**, *AP-49*, 145–149.
14. Ghorbani, K.; Waterhouse, R.B. Ultrabroadband printed (UBP) antenna, IEEE Trans. Antennas Propagation **Dec. 2002**, *AP-50*, 1697–1705.
15. Kumar, G.; Gupta, K.C. Nonradiating edges and four edges gap coupled multiple resonator broadband microstrip antennas. IEEE Trans. Antennas Propagation **Feb. 1985**, *AP-33*, 173–178.
16. Weigand, S.; Huff, G.H.; Pan, K.H.; Bernhard, J.T. Analysis and design of broadband single-layer rectangular U-slot microstrip patch antennas. IEEE Trans. Antennas Propagation **March 2003**, *51*, 457–468.

© 2006 by Taylor & Francis Group, LLC

17. Jackson, D.R.; Williams, J.T.; Bhattacharyya, A.K.; Smith, R.; Buchheit, S.J.; Long, S.A. Microstrip patch designs that do not excite surface waves. IEEE Trans. Antennas Propagation **Aug. 1993**, *41*, 1026–1037.
18. Kraus, J.D. The helical antenna, In *Antennas*; McGraw Hill: NY, 1950; Chap. 7.
19. Adams, A.T.; Greenough, R.F.; Walkenburg, R.K.; Mendelovicz, A.; Lumjiak, C. The quadrifilar helix antenna. IEEE Trans. Antennas Propagation *22* (3), 173–178, 1074.
20. King, H.E.; Wong, J.L. Helical antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; McGraw-Hill: NY, 1993.
21. Elliott, R.S. *Antenna Theory and Design*; Prentice Hall: Englewood-Cliffs, NJ, 1981.
22. Wong, J.L.; King, H.E. Broadband quasi-tapered helical antennas. IEEE Trans. Antennas Propagation **1979**, *27* (1), 72–78.
23. Rumsey, V.H. *Frequency Independent Antennas*; Academic Press: NY, 1966.
24. Corzine, R.G.; Mosko, J.A. *Four Arm Spiral Antennas*; Artech House: Norwood, MA, 1990.
25. Stutzman, W.L.; Thiele, G.A. *Antenna Theory and Design*; Wiley, NY, 1981.
26. DuHamel, R.H.; Scherer, J.P. Frequency independent antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; McGraw-Hill: NY, 1993.
27. Dyson, J.D. The equiangular spiral antenna. IRE Trans. Antennas Propagation **1959**, *7* (2), 181–187.
28. Wang, J.J.H.; Tripp, V.K. Design of multi-octave spiral-mode microstrip antennas. IEEE Trans. Antennas Propagation **1991**, *39* (3), 332–335.
29. Champagne, N.J.; Williams, J.T.; Wilton, D.R. Resistively loaded printed spiral antennas. Electromagnetics **1994**, *14* (3–4), 363–395.
30. Dyson, J.D. The characteristics and design of the conical log-spiral antenna. IEEE Trans. Antennas Propagation **1965**, *13* (7), 488–499.
31. Yeh, Y.S.; Mei, K.K. Theory of conical equiangular-spiral antenna—part II: current distributions and input impedances. IEEE Trans. Antennas Propagation **1968**, *16* (1), 14–21.
32. Walter, C.H. *Traveling Wave Antennas*; McGraw-Hill: New York, 1965.
33. Hessel, A. General characteristics of traveling-wave antennas, In *Antenna Theory* Part 2; Colin, R.E.; Zucher, F.J., Eds.; McGraw-Hill: New York, 1969; Chap 19.
34. Tamir, T. Leaky-wave antennas, In *Antenna Theory*, Part 2; Colin, R.E.; Zucher, F.J. Eds.; McGraw-Hill: New York, Chap 20, 1969.
35. Oliner, A.A. Leaky-wave antennas, In *Antenna Engineering Handbook*; 3rd ed.; Hansen, R.C. Ed.; McGraw-Hill: New York, Chap 10, 1993.
36. Tamir, T.; Oliner, A.A. Guided complex waves, part I: fields at an interface. Proc. Inst. Elec. Eng., Vol. 110, pp. 310–324, Feb. 1963.
37. Tamir, T.; Oliner, A.A.; Guided complex waves, part II: relation to radiation patterns, Proc. Inst. Elec. Eng., vol. 110, pp. 325–334, Feb. 1963.
38. Balanis, C.A. *Antenna Theory*; Wiley: New York, 1997.
39. Goldstone, L.O.; Oliner, A.a. Leaky-wave antennas I: rectangular waveguide. IRE Trans. Antennas Propagation **Oct. 1959**, *AP-7*, 307–319.
40. Harrington, R.F. *Time Harmonic Electromagnetic Fields*; McGraw-Hill: New York, 1963.
41. Guglielmi M.; Boccalone, G. A novel theory for dielectric-inset waveguide leaky-wave antennas, IEEE Trans. Antennas Propagation **April 1991**, *AP-39*, 497–504.
42. Guglielmi, M.; Jackson, D.R. Broadside radiation from periodic leaky-wave antennas. IEEE Trans. Antennas Propagation **Jan. 1993**, *41*, 31–37.
43. Zhao, T.; Jackson, D.R.; Williams, J.T. Radiation characteristics of a 2D periodic slot leaky-wave antenna. IEEE AP-S/URSI Intl. Symp. Digest, pp. 482-485, San Antonio, TX, June 16–21, 2002.
44. Jackson, D.R.; Oliner, A.A. A leaky-wave analysis of the high-gain printed antenna configuration. IEEE Trans. Antennas Propagation **July 1988**, *36*, 905–910.
45. Love, A.W. Ed., *Electromagnetic Horn Antennas*; IEEE Press: NY, 1976.
46. Love, A.W. Horn antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; McGraw-Hill: NY, 1993.

© 2006 by Taylor & Francis Group, LLC

47. Clarricoats, P.J.B.; Poulton, G.T. High-efficiency microwave reflector antennas—a review," Proc. IEEE, Vol. 65, No. 10, pp. 1470–1504, 1977.
48. Kelleher, K.S.; Hyde, G. Reflector antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; McGraw-Hill: NY, 1993.
49. Sharpe, E.D. "A triangular arrangement of planar-array elements that reduces the number needed." IEEE Trans. Antennas Propagation **Mar. 1961**, *AP-9*, 126–129.
50. Silver, S. *Microwave Antenna Theory and Design*, *M.I.T. Radiation Laboratory Series*; McGraw-Hill: New York, 1949; Vol. 12.
51. Skolnik, M.I. *Introduction to Radar Systems*; McGraw-Hill: New York, 2001.
52. Taylor, T.T. Design of line source antennas for narrow beamwidth and low side lobes. IRE Trans. **1955**, *AP-7*, 16–28.
53. Hansen, R.C. *Phased Array Antennas*; Wiley: New York, 1998.
54. Carver, K.R.; Mink, J.W. Microstrip antenna technology. IEEE Trans. Antennas Propagation **Jan. 1981**, *AP-29*, 2–24.
55. Liu, C.; Hessel, A.; Shmoys, J. Performance of probe-fed microstrip-patch element phased arrays. IEEE Trans. Antennas Propagation **Nov. 1988**, *AP-36*, 1501–1509.
56. Pozar, D.M.; Schaubert, D.H. Analysis of an infinite array of rectangular microstrip patches with idealized probe feeds. IEEE Trans. Antennas Propagation **1984**, *AP-32*, 1101–1107.
57. Mailloux, R.J. *Phased Array Antenna Handbook*; Artech House: Norwood, MA, 1994.
58. Dion, A. Nonresonant slotted arrays. IRE Trans. Antennas Propagation **Oct. 1958**, *AP-6*, 360–365.
59. Buck, G.J. Quantization and reflection lobe dispersion, In *Phased Array Antennas*; Oliner, A.A.; Knittel, G.H. Eds.; Artech House: Norwood, MA, 1972.
60. Ajioka, J.S. Frequency scan antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; Mc-Graw Hill: New York, 1993.
61. Wheeler, H.A. Simple relations derived from a phased-array antenna made of an infinite current sheet. IEEE Trans. Antennas Propagation **Jul. 1965**, *AP-13*, 506–514.
62. Pozar, D.M. General relations for a phased array of printed antennas derived from infinite current sheets. IEEE Trans. Antennas Propagation **May 1985**, *AP-33*, 498–504.
63. Wheeler, H.A. The grating-lobe series for the impedance variation in a planar phased-array antenna. IEEE Trans. Antennas Propagation **Nov. 1966**, *AP-14*, 707–714.
64. Hansen, R.C.; Brunner, G. Dipole mutual impedance for design of slot arrays. Microwave J. **Dec. 1979**, *22*, 54–56.
65. Knittel, G.H.; Hessel, A.; Oliner, A.A. Element pattern nulls in phased arrays and their relation to guided waves, Proc. IEEE, Vol. 56, pp. 1822–1836, Nov. 1968.
66. Frazita, R.F. Surface-wave behavior of a phased array analyzed by the grating-lobe series. IEEE Trans. Antennas Propagation **Nov. 1967**, *AP-15*, 823–824.
67. Pozar, D.M.; Schaubert, D.H. Scan blindness in infinite phased arrays of printed dipoles. IEEE Trans. Antennas Propagation. June 1984, *AP-32*, 602–610.
68. Edelberg, S.; Oliner, A.A. Mutual coupling effects in large antenna arrays: part I-slot arrays. IRE Trans. **May 1960**, *AP-8*, 286–297.
69. Tsandoulas, G.N.; Knittel, G.H. The analysis and design of dual-polarization square-waveguide phased arrays. IEEE Trans. Antennas Propagation **Nov. 1973**, *AP-21*, 796–808.
70. Lee, S.W.; Jones, W.R. On the suppression of radiation nulls and broadband impedance matching of rectangular waveguide phased arrays. IEEE Trans. Antennas Propagation **Jan. 1971**, *AP-19*, 41–51.
71. Clavin, A.; Huebner, D.A.; Kilburg, F.J. An improved element for use in array antennas. IEEE Trans. Antennas Propagation **July 1974**, *AP-22*, 521–526.
72. Munk, B.A.; Kornbau, T.W.; Fulton, R.D. Scan independent phased arrays. Radio Sci. **Nov.–Dec. 1979**, *14*, 979–990.

© 2006 by Taylor & Francis Group, LLC

10
Electromagnetic Compatibility

Christos Christopoulos
University of Nottingham,
Nottingham, England

10.1. SIGNIFICANCE OF EMC TO MODERN ENGINEERING PRACTICE

The term *electromagnetic compatibility* (EMC) stands for the branch of engineering dealing with the analysis and design of systems that are compatible with their electromagnetic environment. It may be claimed that there are two kinds of engineers—those who have EMC problems and those who will soon have them. This statement illustrates the impact of EMC on modern engineering practice.

Interference problems are not new. Since the beginning of radio engineers noticed the difficulties encountered when trying to make ground connections to the chassis of different systems and the onset of whistling noise attributed to atmospheric conditions. All these are manifestations of electromagnetic interference (EMI) and demonstrate the need to design systems which are compatible with their electromagnetic environment.

There are two aspects to EMC. First, systems must be designed so that they do not emit significant amounts of unintended electromagnetic (EM) radiation into their environment. This aspect is described as *emission* and may be divided in turn into *conducted* and *radiated* emission. Second, systems must be capable of operating without malfunction in their intended environment. This aspect is described as *immunity*, or alternatively, as *susceptibility*. Hence, all EMC analysis and design techniques aim to address these two aspects using circuit-based and field-based experimental, analytical, and numerical techniques.

It is important to realize why EMC has become so important in recent years. As is usual in such cases, there are several reasons:

Modern design relies increasingly on the processing of digital signals, i.e., signals of a trapezoidal shape with very short rise and fall times. This gives them a very broad frequency spectrum and thus they are more likely to interfere with other systems.

Most modern designs rely on clocked circuits with clock frequencies exceeding 2 GHz. This implies very short transition times (see above) and also the presence of several harmonics well into the microwave region. Such a broad spectrum makes it inevitable that some system resonances will be excited forming efficient antennas for radiating EM energy into the environment and coupling to other systems.

© 2006 by Taylor & Francis Group, LLC

Voltage levels for switching operations have steadily decreased over the years from hundreds of volts (vacuum tubes) to a few volts in modern solid-state devices. This makes systems more susceptible to even small levels of interference.

We make a much greater use of the EM spectrum as, for instance, with mobile phones and other communication services.

Equipment is increasingly constructed using small cabinets made out of various plastics and composites in contrast to traditional design, which used metal (a good conductor) as the primary constructional material. This trend meets the need for lighter, cheaper, and more aesthetically pleasing products. However, poor conductors are not good shields for EM signals, thus exacerbating emission and susceptibility problems.

Miniaturization is the order of the day, as smaller, lighter mobile systems are required. This means close proximity between circuits and thus greater risk of intrasystem interference (cross talk).

We rely increasingly on electronics to implement safety critical functions. Examples are, antilock break systems for cars, fly-by-wire aircraft, etc. It is, therefore, imperative that such circuits be substantially immune to EMI and hence malfunction.

We might add here military systems that use electronics substantially and are continuously exposed to very hostile EM environments either naturally occurring (e.g., lightning) or by deliberate enemy action (e.g., jamming).

These points illustrate the engineering need to design electromagnetically compatible systems. International standardization bodies have recognized for many years the need to define standards and procedures for the certification of systems meeting EMC requirements. The technical advances outlined above have given a new impetus to this work and have seen the introduction of international EMC standards covering most aspects of interference control and design. These are the responsibility of various national standard bodies and are overseen by the International Electrotechnical Commission (IEC) [1].

The impact of EMC is thus multifaceted. The existence of EMC design procedures which adhere to international standards, ensures that goods may be freely moved between states and customers have a reasonable expectation of a well engineered, reliable and safe product. However, meeting EMC specifications is not cost free. The designer needs to understand how electromagnetic interactions affect performance, and implement cost effective remedies. A major difficulty in doing this is the inherent complexity of EM phenomena and the lack of suitably qualified personnel to do this work. This is a consequence of the fact that for several decades most engineers focused on digital design and software developments with little exposure to EM concepts and radio-frequency (RF) design. In this chapter we aim to describe how EM concepts impact on practical design for EMC and thus assist engineers wishing to work in this exciting area. It is also pointed out that modern high-speed electronics have to cope in addition to EMC also with signal integrity (SI) issues. The latter is primarily concerned with the propagation of fast signals in the compact nonuniform environment of a typical multilayer printed-circuit board (PCB). At high clock rates the distinction between EMC and SI issues is somewhat tenuous as the two are intricately connected. Thus most material presented in this chapter is also relevant to SI.

We emphasize predictive EMC techniques rather than routine testing and certification as the art in EMC is to ensure, by proper design, that systems will meet

© 2006 by Taylor & Francis Group, LLC

specifications without the need for extensive reengineering and modification. It is in this area that electromagnetics has a major impact to make. It is estimated that up to10% of the cost of a new design is related to EMC issues. This proportion can be considerably higher if proper EM design for EMC has not been considered at the start of the design process. The interested reader can access a number of more extensive books on EMC and SI. The EMC topic is also taken up in my own *Principles and Techniques of Electromagnetic Compatibility* [2]. A general text on SI is Ref. 3. Other references are given in the following sections.

We start in the following section with a brief survey of useful concepts from EM field theory, circuits, and signals as are adapted for use in EMC studies. There follow sections on coupling mechanisms, practical engineering remedies to control EMI and EMC standards and testing. We conclude with an introduction of some new concepts and problems which are set to dominate EMC studies in the years to come.

10.2. USEFUL CONCEPTS AND TECHNIQUES FROM ELECTROMAGNETICS, SIGNALS, AND CIRCUITS

In this section we summarize useful concepts for EMC. Most readers will be familiar with this material but may find it still useful as it is presented in a way that is useful to the EMC engineer.

10.2.1. Elements of EM Field Theory

Most EMC standards and specifications are expressed in terms of the electric field. There are cases where the magnetic field is the primary consideration (e.g., shielding at low frequencies) but these are the minority. In emission studies, the electric field strength is specified at a certain distance from the equipment under test (EUT). These distances are typically, 1 m (for some military specifications), 3 m, 10 m, and 30 m. Measurements or calculations at one distance are then extrapolated to estimate the field at another distance, assuming far-field conditions. This implies an extrapolation law of $1/r$, where r is the distance. This is only accurate if true far-field conditions are established and this can only be guaranteed if the extrapolation is done from estimates of the field taken at least a wavelength away from the EUT. This is not always the case, but the practice is still followed, thus introducing considerable errors in field estimates.

In EMC work electric fields are normally expressed in decibels relative to some reference. A commonly employed reference is 1 μV/m. Thus an electric field E in V/m can be expressed in dBμV/m

$$E \text{ dB}\mu\text{V/m} = 20 \log\left(\frac{E}{1 \times 10^{-6}}\right) \tag{10.1}$$

Thus, an electric field of 10 mV/m is equal to 80 dBμV/m. Typical emission limits specified in various standards range between 30 and 55 dBμV/m. Similar principles apply when the magnetic field H is expressed normally to a reference of 1 μA/m.

A lot of reliance is placed in EMC analysis on quasistatic concepts. This is due to the desire of designers to stay with familiar circuit concepts and also to the undeniable complexity of working with EM fields at high frequencies. Strictly speaking, quasistatic

© 2006 by Taylor & Francis Group, LLC

concepts apply when the physical size of the system D is much smaller than the shortest wavelength of interest $D \ll \lambda$. This is often the case but care must be taken before automatic and indiscriminate use of this assumption is made. Assuming that the quasi-static assumption is valid, we can then talk about the capacitance and inductance of systems and have a ready-made approach for the calculation of their values. Important in many EMC calculations is therefore the extraction of the L and C parameters of systems so that a circuit analysis can follow. This is, in general, much simpler than a full-field analysis and is to be preferred provided accuracy does not suffer.

There are many ways to extract parameters using a variety of computational electromagnetic (CEM) techniques. Whenever an analytical solution is not available [4], CEM techniques such as the finite element method (FEM), the method of moments (MoM), finite-difference time-domain (FDTD) method, and the transmission-line modeling (TLM) method may be employed [5–8]. All such calculations proceed as follows.

A model of the system is established normally in two-dimensions (2D) to obtain the per unit length capacitance. The systems is electrically charged and the resulting electric field is then obtained. The voltage difference is calculated by integration and the capacitance is then finally obtained by dividing charge by the voltage difference. If for instance the parameters of a microstrip line are required, two calculations of the capacitance are done. First with the substrate present and then with the substrate replaced by air. The second calculation is used to obtain the inductance from the formula $L = 1/(c^2 C_0)$, where c is the speed of light ($= 3 \times 10^8$ m/s) and C_0 is the capacitance obtained with the substrate replaced by air. This approach is justified by the fact that the substrate does not normally affect magnetic properties. If this is not the case (the substrate has relative magnetic permeability other than one), then a separate calculation for L must be done by injecting current I into the system, calculating the magnetic flux Φ linked, and thus the inductance $L = \Phi/I$. It is emphasized again that when quasistatic conditions do not apply, the concept of capacitance is problematic as the calculation of voltage is not unique (depends on the path of integration). Similar considerations apply to inductance. At high frequencies, therefore, where the wavelength gets comparable with the size of systems, full-field solutions are normally necessary. This increases complexity and requires sophisticated modeling and computational capabilities. The reader is referred to Ref. 2 for a more complete discussion of the relationship between circuit and field concepts.

In EMC work it is important to grasp that what is crucial is not so much the visible circuit but stray, parasitic, components. This is where an appreciation of field concepts can assist in interpretation and estimation of relevant parameters and interactions. The reason that parasitic components are so important is that they affect significantly the flow of common mode currents. This is explained in more detail further on in this section. Particular difficulties in EMC studies are encountered at high frequencies. Here the quasistatic approximation fails and full-field concepts must be employed. At high frequencies, fields are generally not guided by conductors and spread out over considerable distances. Before we focus on high-frequency problems we state more clearly the range of applicability of the various models used to understand electrical phenomena. Generally, electrical problems fit into three regimes:

1. When the size of a system is smaller than a wavelength in all three dimensions, then it may be adequately represented by lumped component equivalent circuits. Solution techniques are those used in circuit analysis. This is the simplest case, and it is preferred whenever possible.

© 2006 by Taylor & Francis Group, LLC

2. When a system is smaller than a wavelength in two dimensions and comparable or larger to a wavelength in the third dimension, then the techniques of transmission-line analysis can be used. These are based on distributed parameter equivalent circuits.
3. When a system is electrically large in all three dimensions, then full-field calculations must be employed based on the full set of Maxwell's equations.

Clearly, the last case offers the most general solution, and it is the most complex to deal with. In this case it is normally necessary to employ numerical techniques such as those described in [5–8].

We focus here on some of the most useful EM concepts that are necessary to understand the high-frequency behavior of systems. At high frequencies EM energy is transported in a wavelike manner. This is done either in the form of guided waves as in a transmission line or in free space as from a radiating antenna. Taking for simplicity the case of wave propagation in one dimension z, then the electric field has only a y component and the magnetic field an x component. The electric field behavior is described by the wave equation

$$\frac{\partial^2 E_y}{\partial x^2} = \frac{1}{u^2}\frac{\partial^2 E_y}{\partial t^2} \tag{10.2}$$

where, u is the velocity of propagation in the medium concerned,

$$u = \frac{1}{\sqrt{\mu\varepsilon}} \tag{10.3}$$

In the case of propagation in free space, u is equal to the speed of light. An identical equation describes magnetic field behavior. Transport of EM energy after a few wavelengths away from radiating structures, such as the various interconnects, wiring, etc., in electrical systems takes place in accordance to Eq. (10.2). In the so-called far field E and H are transverse to each other and their magnitudes are related by the expression,

$$H = \frac{E}{\eta} \tag{10.4}$$

where, η is the intrinsic impedance of the medium. In the case of free space

$$\eta = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 377\,\Omega \tag{10.5}$$

In EMC calculations it is customary to calculate the magnetic field from the electric field using Eqs. (10.4) and (10.5). This is however only accurate if plane wave conditions apply and this is generally true at a distance exceeding approximately a wavelength away from the radiator. In the near field, the field retains some of the character of the radiating structure that produced it. If the radiator is in the form of a dipole, where voltage differences are accentuated, then the electric field is higher than would be expected for plane wave conditions and the wave impedance is larger than 377 Ω. If however, the radiating structure is in the form of a loop, where currents are accentuated,

© 2006 by Taylor & Francis Group, LLC

then the impedance of the medium is smaller than 377 Ω, and the magnetic field predominates. In either case, in the far field the wave impedance settles at 377 Ω.

For a short dipole, the magnitude of the wave impedance as a function of the distance r away from it is given by the formula

$$|Z_w| = \eta \frac{\sqrt{1 + 1/(\beta r)^6}}{1 + 1/(\beta r)^2} \tag{10.6}$$

where, $\beta r = 2\pi r/\lambda$. It is clear that as $r \gg \lambda$, the wave impedance tends to η.

As an example, we give here the formulas for the field near a very short (Hertzian) dipole.

The configuration is shown in Fig. 10.1 with the components in spherical coordinates.

$$\begin{aligned}
E_\vartheta &= \frac{j\omega\mu}{4\pi}(I\Delta l)\frac{e^{-j\beta r}}{r}\sin\vartheta\left[1 + \frac{1}{j\beta r} + \frac{1}{(j\beta r)^2}\right] \\
E_r &= \frac{j\omega\mu}{2\pi}(I\Delta l)\frac{e^{-j\beta r}}{r}\cos\vartheta\left[\frac{1}{j\beta r} + \frac{1}{(j\beta r)^2}\right] \\
E_\varphi &= 0 \\
H_r &= H_\vartheta = 0 \\
H_\varphi &= \frac{(I\Delta l)}{4\pi}\frac{e^{-j\beta r}}{r}\sin\vartheta\left[j\beta + \frac{1}{r}\right]
\end{aligned} \tag{10.7}$$

where, $\beta = 2\pi/\lambda$ is the phase constant, I is the current, and Δl is the length of the short dipole. It is clear from these formulas that the field varies with the distance r from the dipole in a complex manner. This is particularly true when r is small (near field) when all terms in the right-hand side of Eq. (10.7) are of significant magnitude. In the far field ($r \gg \lambda$), the field simplifies significantly,

$$\begin{aligned}
E_\vartheta &\simeq j\eta\frac{\beta(I\Delta l)}{4\pi}\frac{e^{-j\beta r}}{r}\sin\vartheta \\
H_\varphi &\simeq j\frac{\beta(I\Delta l)}{4\pi}\frac{e^{-j\beta r}}{r}\sin\vartheta
\end{aligned} \tag{10.8}$$

We notice that in far field only two components of the field remain which are orthogonal to each other, and they both decay as $1/r$. This is characteristic of a radiation field.

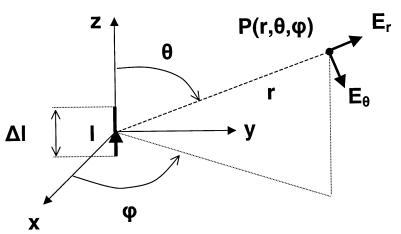


Figure 10.1 Coordinates used for calculating the field components of a very short dipole.

© 2006 by Taylor & Francis Group, LLC

In complex systems with numerous radiating wire segments field behavior is very complex and it can only be studied in detail with powerful modeling tools. Similar formulas apply for short loops. Formulas for radiation from antennas may be found in Ref. 9 and other similar texts on antenna theory.

10.2.2. Treatment of Signals and Sources

The study and characterization of the EMC behavior of systems require an understanding of the nature of electrical signals encountered in engineering practice. One can classify signals in several ways depending on the criterion selected.

Many signals employed during normal engineering work are deterministic in nature, that is, their evolution in time can be precisely predicted. However, in many cases of signals with noise, we cannot predict precisely their time evolution. We call these signals random or stochastic. We can however make precise statements about them which are true in the statistical sense. The study of random signals requires sophisticated tools which are beyond the scope of this chapter. For a brief introduction see Ref. 2 and for a fuller treatment see Ref. 10. We will limit our discussion here to deterministic signals.

Some signals consist of essentially a single frequency (monochromatic or narrow-band). A signal that occupies a very narrow band in the frequency spectrum, persists for a long period in time. A typical example is a steady-state sinusoidal signal. Other signals occupy a wide band of frequencies and therefore persist for relatively short periods in time. Typical examples are pulses of the kind found in digital circuits.

Whatever the nature of the signal, we can represent it as the weighted sum of a number of basis functions. A very popular choice of basis functions are harmonic functions, leading to representation of signals in terms of Fourier components [11]. For a periodic signal we obtain a Fourier series and for an aperiodic signal a Fourier transform. As an example, we give the Fourier series components of a signal of great engineering importance—the pulse train shown in Fig.10.2. For this signal the period is T, the duty cycle is τ/T and the transition time (rise and fall time) is τ_r. This trapezoidal-shaped pulse is a good representation of pulses used in digital circuits. The Fourier spectrum of this signal is given below:

$$|A_n| = 2V_0 \frac{\tau - \tau_r}{T} \left| \frac{\sin[\pi n(\tau - \tau_r)/T]}{\pi n(\tau - \tau_r)/T} \right| \left| \frac{\sin(\pi n \tau_r/T)}{\pi n \tau_r/T} \right| \tag{10.9}$$

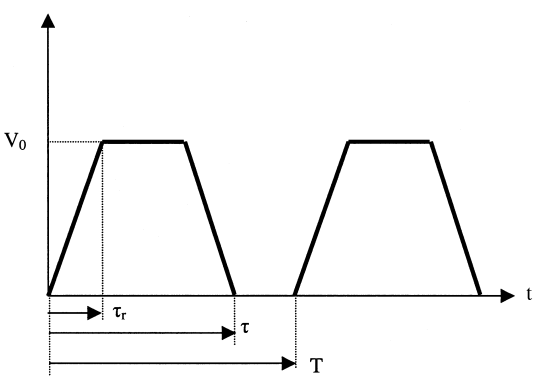


Figure 10.2 Typical trapezoidal pulse waveform.

© 2006 by Taylor & Francis Group, LLC

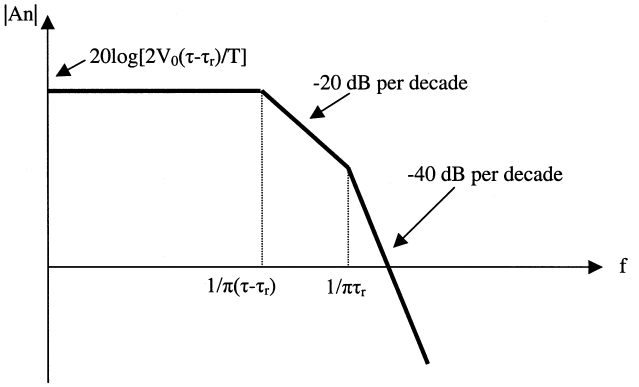


Figure 10.3 Envelope of the amplitude spectrum of the waveform in Fig. 10.2.

where, $n = 0, 1, 2, \ldots,$ and $A_0 = 2V_0(\tau - \tau_r)/T$. Equation (10.9) represents a spectrum of frequencies, all multiples if $1/T$, with amplitudes which are modulated by the $(\sin x)/x$ functions. Three terms may be distinguished: a constant term $2V_0(\tau - \tau_r)/T$ independent of frequency, a term of magnitude 1 up to frequency $1/[\pi(\tau - \tau_r)]$ thereafter decreasing by 20 dB per decade of frequency, and term of magnitude 1 up to frequency $1/\pi\tau_r$ thereafter decreasing by 20 dB per decade. The envelope of the amplitude spectrum for a trapezoidal pulse train is shown in Fig. 10.3. The shorter the transition time, the higher the frequency at which the amplitude spectrum starts to decline. Short rise times imply a very wide spectrum of frequencies.

It is customary in EMC to study the behavior of systems as a function of frequency. However, increasingly, other techniques are used to speed up experimentation and analysis where a system is excited by short pulses. The former case is referred to as analysis in the frequency domain (FD) and the latter as analysis in the time domain (TD). The two domains are related by the Fourier transform as explained further in the next subsection.

Commonly encountered sources of EMI are characterized as far as possible using standard signal waveforms [2]. Amongst naturally occuring EMI sources most prominent is lightning [12,13] because of its wide spectrum and wide geographical coverage. A general background noise level due to a variety of cosmic sources exists, details of which may be found in Ref. 14. There is also a range of man-made sources including radio transmitters [15,16], electroheat equipment [17], digital circuits and equipment of all kinds [18], switched-mode power supplies and electronic drives [19,20], electrostatic discharge [21], and for military systems NEMP [22,23]. A survey of general background levels of man-made noise may be found in Ref. 24. Reference 25 describes the methodology to be used to establish the nature and severity of the EM environment on any particular site.

10.2.3. Circuit Analysis for EMC

As already mentioned lumped circuit component representation of systems and hence circuit analysis techniques are used whenever possible in EMC. For the serious student of EMC familiarity with the relationship of circuit and field concepts is very useful. As soon as it has been established that a circuit representation of a system is adequate, normal circuit analysis techniques may be employed [26]. In general circuits can be studied in two ways.

© 2006 by Taylor & Francis Group, LLC

First, the frequency response may be obtained. The source signal is analysed into its Fourier components $V_{in}(j\omega)$, and the output is then obtained from the frequency response or transfer function $H(j\omega)$ of the circuit,

$$V_{out}(j\omega) = H(j\omega)V_{in}(j\omega) \tag{10.10}$$

Full-field analysis in the FD is based on the same principles, but the transfer function is much more complex and often cannot be formulated in a closed form.

Second, the problem may be formulated in the time domain whereby the system is characterized by its response to an impulse, by the so-called *impulse response* $h(t)$. The response to any source signal $v(t)$ is then given by the convolution integral,

$$v_{out}(t) = \int_{\infty}^{\infty} v_{in}(\tau)h(t-\tau)\,d\tau \tag{10.11}$$

Full-field analysis in the TD is done in the same way but the impulse response is a much more complex function which often cannot be formulated in a closed form. In linear systems Eqs. (10.10) and (10.11) are equivalent formulations as the two response functions $H(j\omega)$ and $h(t)$ are Fourier transform pairs. However, in nonlinear systems, where the principle of superposition does not apply, only the time domain approach can be employed. Full-field solvers broadly reflect these limitations.

Simple nonlinear circuits are used in EMC to implement various detector functions (e.g., peak and quasi-peak detectors). For a discussion of detector functions, see Refs. 27 and 28.

10.3. IMPORTANT COUPLING MECHANISMS IN EMC

In every EMC problem we may distinguish three parts as shown in Fig. 10.4. These are the source of EMI, the victim of EMI and a coupling path. If at least one of these three parts is missing then we do not have an EMC problem. In the previous section we have discussed some of the sources and circuits which may be victims to interference. In the present section we focus on the coupling mechanisms responsible for EMI breaching the gap between source and victim. A comprehensive treatment of this extensive subject is beyond the scope of this chapter. The interested reader is referred to comprehensive texts on EMC such as [2,29–31]. We will however present here the essential principles of EM coupling.

10.3.1. Penetration Through Materials

In many systems the outer skin (e.g., aircraft) or enclosure (e.g., equipment cabinet) forms part of an EM shield which contributes to the reduction of emission and susceptibility

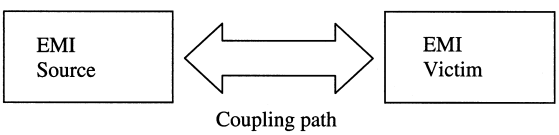


Figure 10.4 Source, coupling path, and victim of EMI.

© 2006 by Taylor & Francis Group, LLC

problems. A perfectly conducting shield without apertures or penetrations would be an ideal shield for all but low-frequency magnetic fields. However such an ideal is difficult to approach in practice. Invariably, shields are not perfectly conducting and have several openings and through wire connections. In this subsection, we focus on penetration through the walls of a shield due to its finite electrical conductivity.

In this and other shielding problems it is important to use the concept of *shielding effectiveness* (SE). SE is defined as the ratio in dB of the field without and with the shield.

$$\mathrm{SE} = 20\log\left|\frac{E_0}{E_t}\right| \tag{10.12}$$

A similar expression is used for the magnetic shielding effectiveness.

The SE of canonical shapes such as spheres, cylinders made out of various materials may be calculated analytically. Of particular relevance in practical applications is the SE due to the material itself at low frequencies and particularly to the magnetic field. Taking as an example a very long cylinder of inner radius a and wall thickness D, the ratio of incident to transmitted longitudinal magnetic field (low-frequency, displacement current neglected) is given by the formula [2,32–34],

$$\frac{H_i}{H_t} = \cosh\gamma D + \frac{\gamma a}{2\mu_r}\sinh\gamma D \tag{10.13}$$

where, γ is the propagation constant inside the wall material $\gamma = (1+j)/\delta$ and δ is the skin depth. The skin depth is given by the formula,

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} \tag{10.14}$$

In this expression μ is the magnetic permeability of the wall material and σ is the electrical conductivity. For such configurations, shielding for both electric and magnetic fields can be understood by the two simple equivalent circuits shown in Fig. 10.5. For a thin-walled spherical cell and low frequencies the parameters shown in Fig. 10.5 are given approximately by, $C = 3\varepsilon_0 a/2$, $L = \mu_0 a/3$, $R = 1/\sigma D$. Study of this circuit gives

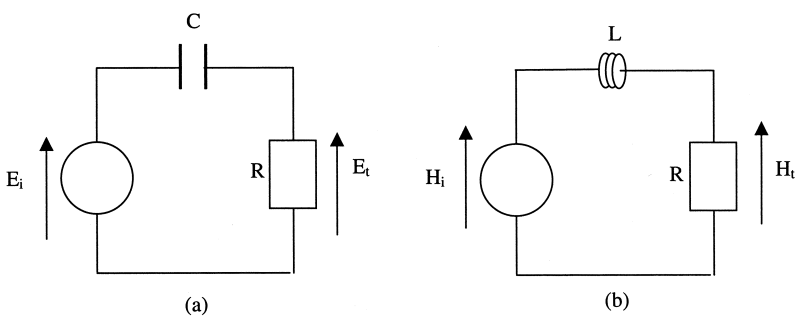


Figure 10.5 Circuit analogs for SE (a) for electric and (b) for magnetic fields.

© 2006 by Taylor & Francis Group, LLC

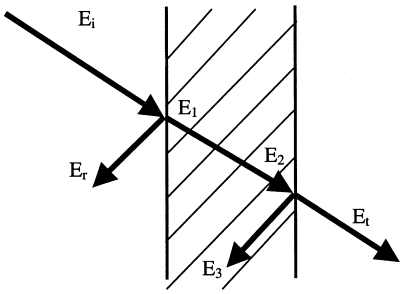


Figure 10.6 Wave approach to penetration through walls.

a good insight into some of the problems encountered with shielding. In each case subscript i indicates the incident field and t the transmitted field inside the structure. Hence a high voltage across R in the equivalent circuits indicates poor shielding. Examining electric field shielding first, we observe that at low frequencies (LF) shielding is very good (C has a very high impedance at LF). Hence it is relatively easy to shield against LF electric fields. In contrast, shielding of magnetic field at LF is very difficult (L has a very low impedance at LF). The shielding of LF magnetic fields requires special arrangements based on forcing it to divert into very high permeability (low reluctance) paths. As a general comment, reductions in R improve shielding. Hence any slots and/or obstructions on the surface of the shield must be placed in such a way that they do not obstruct the flow of eddy currents (thus keeping R low). Further formulas for diffusive penetration through shields for some canonical shapes may be found in Ref. 35.

Another approach to diffusive shielding is based on the wave approach [2,36,37]. This approach is depicted in Fig. 10.6 where an incident electric field E_i is partially reflected from the wall (E_r), partially penetrates (E_1), reaches the other side of the wall after some attenuation (E_2), suffers a partial internal reflection (E_3), and part of it is transmitted into the inner region (E_t). Component E_3 suffers further reflections (not shown) which contribute further to the transmitted wave. In complex problems, numerical solutions are necessary which employ special thin-wall formulations which also allow for inhomogenuities and anisotropies [38,39].

It should be emphasized that although we have discussed shielding here by illustrating penetration from an outer region to an inner region, the reverse process follows the same rules (equivalence principle).

10.3.2. Penetration Through Apertures

A major route for penetration of EM radiation is through apertures. By this we mean any hole, opening, ventilation grid, imperfect joint which breaches the continuity of the conducting shield. It is normally the case that apertures form the major route for radiation breaching a shield. Aperture penetration may be tackled in different ways depending on circumstances. These are based on small hole theory, simple analytical formulations for slots, intermediate level tools, and full numerical models. We examine each approach below.

1. For holes that are electrically small we first calculate the electric field E_{sc} at the position of the whole assuming that the aperture has been replaced by a

© 2006 by Taylor & Francis Group, LLC

perfect conductor (short-circuit electric field). The presence of the aperture is then represented by placing an equivalent dipole inside the wall, where the aperture is again replaced by a perfect conductor. The dipole moment of the dipole is [40]

$$p_e = 2\varepsilon\alpha_e E_{\text{sc}} \tag{10.15}$$

where α_e is the hole electric polarizability [41]. As an example, the polarizability of a round hole of diameter d is $\alpha_e = d^3/12$. The inner field can then be obtained by using antenna theory or any other suitable technique.

2. Alternative formulations have appeared in the literature where calculations of shielding effectiveness have been made for simple commonly encountered apertures. Particularly well known is the SE of a slot of length ℓ [29]:

$$\text{SE} = 20\log\frac{\lambda}{2\ell} \tag{10.16}$$

If the length of the slot is 1/10 of the wavelength then SE = 14 dB. Such performance at 1 GHz implies slot lengths smaller than 3 cm. Clearly the shorter the length the higher the SE. For the same area of aperture it is better to have several smaller apertures rather than one large one. The formula above for N apertures modifies to

$$\text{SE} = 20\log\frac{\lambda}{2\ell\sqrt{N}} \tag{10.17}$$

Equations (10.16) and (10.17) do not take into account either the width of the slot or the presence of a resonant equipment enclosure hence they may result in large errors in SE estimates.

3. Intermediate level tools can make good estimates of SE with a minimum of computational effort and are thus a compromise between accuracy and computational efficiency. The basic configuration is given in Fig. 10.7a and the intermediate level model in Fig. 10.7b [42–45]. The model of penetration through the aperture and propagation in the cabinet is broken down to three components:

   a. First, the incident field is represented by a simple Thevenin equivalent circuit, where the impedance is the intrinsic impedance of free space.
   b. Second, the aperture is represented by two halves of a coplanar strip line, shorted at both ends [46].
   c. Third, the cabinet is represented by a shorted waveguide with an impedance and propagation constant that take account of the first resonant mode.

   The three models are combined to form the complete model shown in Fig. 10.7b. This is relatively simple model to manipulate. The SE for the electric field is simply given in terms of the equivalent circuit parameters,

$$\text{SE} = 20\log\frac{V_0}{2V(z)} \tag{10.18}$$

© 2006 by Taylor & Francis Group, LLC

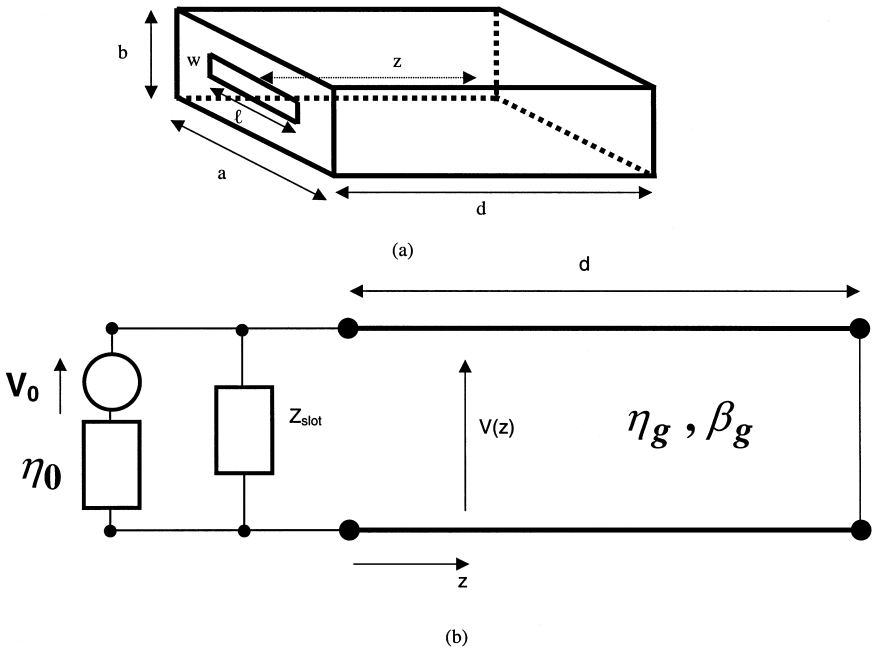


Figure 10.7 Intermediate level model (b) for the SE of a cabinet (a).

SE for the magnetic field is similarly obtained by replacing in Eq. (10.18) voltage by current. Typical results are shown in Fig. 10.8 and they illustrate several important points. At some frequencies, corresponding to cabinet resonances, the SE is negative implying that the presence of the cabinet results in field enhancement. The presence of the cabinet is of major significance in the calculation of SE. The SE has a different value depending on the point chosen to calculate it. Even away from resonances, the simple formula Eq. (10.16) is in considerable error. The introduction of PCBs and other loads inside the cabinet affects SE primarily near resonances. The method of including contents is explained in detail in Refs. 44 and 45. Application of these formulations in industrial problems may be found in Ref. 47.

4. The cabinet and its apertures may be described using one of the full-field solvers described in Refs. 5–8. For the case of a small number of electrically large apertures this process is straightforward [48]. However, in the case of complex and extensive ventilation grids the computational effort required in describing and meshing a large three-dimensional problem is excessive. In such cases, techniques have been developed to calculate SE using full-field models with embedded digital signal algorithms describing the grid of apertures [49–50]. Full-field calculation of SE in densely loaded cabinets, with several apertures, is still a very demanding computational task.

10.3.3. Conducted Penetrations

Conducted penetrations are another major means of introducing EMI into systems. Conducted penetrations may consist of power, control and communication cables which

© 2006 by Taylor & Francis Group, LLC

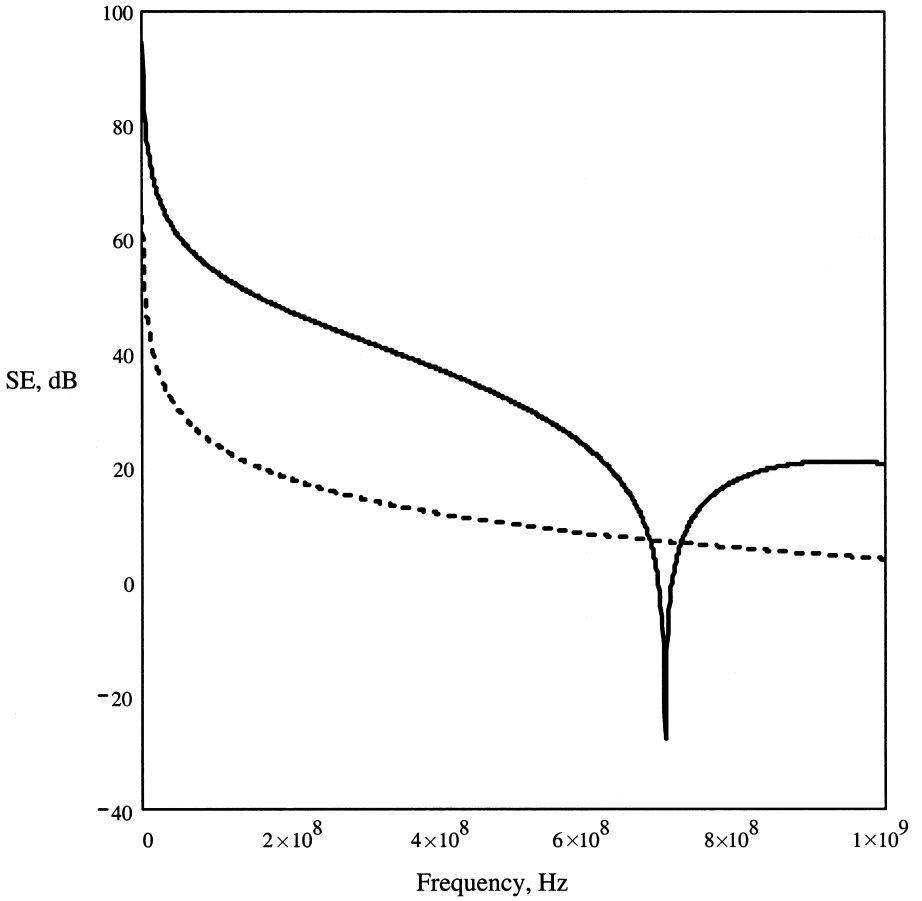


Figure 10.8 Electric field SE of a cabinet ($0.3 \times 0.12 \times 0.3\,\text{m}^3$, slot $0.1\,\text{m} \times 5\,\text{mm}$, $z = 0.15\,\text{m}$). Intermediate model (solid curve), Eq. (10.16) (broken curve).

may be shielded or unshielded. In addition, conducting pipes used for bringing services (water, air, etc.) into buildings and equipment form another route for EMI. Due to the variety of configurations it is difficult to offer general advice and general-purpose models for estimating the level of interference and thus ensuring EMC. We show in Fig. 10.9 in schematic form a penetration of a conductor through a barrier wall without a dc connection between the conductor and the conducting wall [2]. The approach to modeling this penetration is as follows:

We first estimate the coupling of the external field in the portion of the conductor which is the outer region. This can be conveniently done by using antenna theory and working out the coupling of the field to a monopole antenna (conductor above wall) [51,52]. This coupling is represented by the equivalent antenna components V_a and R_a. At the point of entry through the wall we introduce the barrier capacitance C_b to represent high-frequency displacement currents flowing between the conductor and the wall. C_b may be estimated or calculated from a full-field model of the region around the penetration. In the inner region we assume that the conductor is terminated by an equivalent resistance R (or impedance if appropriate), representing in the case of a terminated conductor the

© 2006 by Taylor & Francis Group, LLC

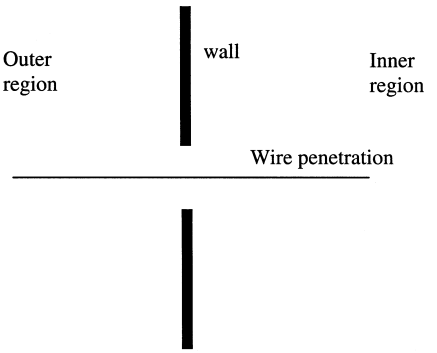


Figure 10.9 Wire penetration through a wall.

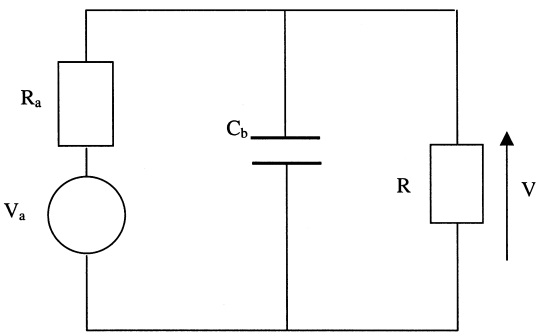


Figure 10.10 Circuit model of penetration for configuration in Fig. 10.9.

actual resistance of the termination, or, in the case of a floating conductor its radiation resistance. The complete approximate circuit is shown in Fig. 10.10. From this circuit we can calculate the voltage across R when the capacitor is present and when it is absent,

$$V_{\text{with } C} = \frac{V_{\text{without } C}}{1 + j\omega RC_b/(1 + R/R_a)} \tag{10.19}$$

From this equation it is clear that at high frequencies the barrier capacitance affords a degree of shielding. A more elaborate arrangement is shown in Fig. 10.11 where a feed through capacitance is shown. This example illustrates the need to use the appropriate model in the prediction of SE. At high frequencies it is not appropriate to use a lumped barrier capacitance as in Fig. 10.10. Taking as a measure of effectiveness the ratio V/I in Fig. 10.11 and treating the feed through capacitor as a short transmission line of length l and characteristic impedance Z_0 we obtain,

$$\frac{V}{I} = \frac{Z_0}{j \sin(2\pi\ell/\lambda)} \tag{10.20}$$

At low frequencies, the impedance in Eq. (10.20) reduces to the impedance of the barrier capacitance. However, when the length of the feed through capacitor approaches

© 2006 by Taylor & Francis Group, LLC

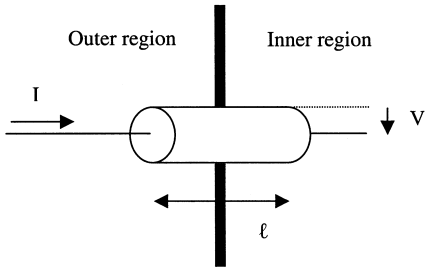


Figure 10.11 A coaxial wire penetration.

the wavelength, the impedance can have very large values. When the length is equal to half the wavelength, the impedance tends to infinity. This illustrates the care that must be taken when constructing models to estimate EMI. Another illustration of this problem is the model required when the penetration in Fig. 10.9 is modified by connecting the conductor to the wall at the entry point using a short length of wire ("pigtail" connection). In such a case it is essential to include in the model the inductance of the pigtail. This then makes it clear that at high frequencies, where the inductive impedance of the pigtail is large, the effectiveness of the connection to the wall is severely reduced. Matters can be improved if a 360° connection of the conductor to the wall is made. In all the cases illustrated above the calculation of the appropriate parameters to include in computations is not a simple matter. Although estimates can normally be made, a full characterization requires full-field EM calculations and the extraction from these of the required parameters. Further discussion of the treatment of wire penetration may be found in Refs. 52 and 53.

An important aspect of EMC analysis and design is the propensity of cables, which are used extensively as interconnects, to pick up and emit EM radiation. Cables with braided shields do not afford complete protection—a certain amount of radiation penetrates. This is traditionally described in terms of a transfer impedance relating the electric field parallel to the inner surface of the shield to the current flowing in the outer surface,

$$Z_T = \frac{E}{I} \tag{10.21}$$

For solid shields of thickness D and inner radius a, the transfer impedance can be calculated analytically [2] and is given by

$$\frac{|Z_T|}{R_{\text{dc}}} \simeq \begin{cases} 1 & D \ll \delta \\ 2\sqrt{2}\dfrac{D}{\delta} e^{-D/\delta} & D \gg \delta \end{cases} \tag{10.22}$$

where, δ is the skin depth, and $R_{\text{dc}} = 1/(2\pi a \sigma D)$ is the dc resistance of the shield.

The situation is much more complex for braided shields. Here, account must be taken of the small holes between strands, incomplete contact between strands, differences in spacing, etc. [54,55]. This is done by adding to a modified Z_T an additional reactive term to account for magnetic field coupling through holes and other imperfections in the shield

© 2006 by Taylor & Francis Group, LLC

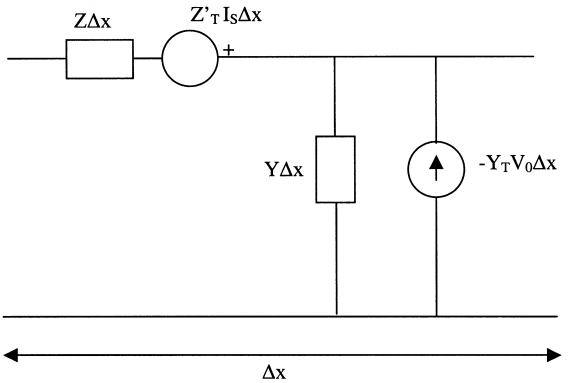


Figure 10.12 Model of a cable segment including transfer impedance and admittance.

Table 10.1 Magnitude of Cable Transfer Impedance (Typical Values in mΩ/m)

Cable type	0.1 MHz	1 MHz	10 MHz	100 MHz
URM102	5	3	4	23
URM43	11	25	158	1585
UR91	2	1	2.5	14
UR67	6	12	55	142
RG62	9	17	100	
RG228	3	0.5	1	
RG22	1	0.04	0.06	

$$Z'_T = Z_T + j\omega M' \tag{10.23}$$

In a similar manner, electric field penetration due to coupling through the capacitance to the inner conductor may be accounted for by a transfer admittance

$$Y_T = j\omega C' \tag{10.24}$$

Further details may be found in Refs. 56 and 57. An equivalent circuit of propagation in a cable taking into account coupling through the shield is shown in Fig. 10.12, where the voltage source accounts for magnetic filed coupling (normally the most significant coupling term), and the current source for electric field coupling (negligible in a well-constructed shield). Some typical values of transfer impedance for commercially available cables are shown in Table 10.1.

10.3.4. Radiation and Cross Talk

An important consideration affecting both EMC and SI is the coupling between adjacent circuits which are in the near field of each other (cross talk), and coupling over large distances through radiation either in the form of emission from circuits, or in the form of coupling of external fields onto circuits. First we tackle near-field coupling (cross talk), and then we examine far-field radiative coupling.

© 2006 by Taylor & Francis Group, LLC

At low frequencies, coupling in the near field can be understood in terms of mutual capacitance and inductance between circuits. A simple approach to this problem is shown in Fig. 10.13. We limit the treatment to static, low-frequency solutions. Capacitive coupling in the case of the configuration in Fig. 10.13a gives the voltage induced on the second conductor due to a voltage on the first conductor as

$$V_2 = \frac{C_{12}}{C_{12} + C_2} V_1 \tag{10.25}$$

where the capacitance in this expression are as marked in the figure. If the second conductor is shielded, but the shield is not grounded, the shield potential will be

$$V_s = \frac{C_{1s}}{C_{1s} + C_s} V_1 \tag{10.26}$$

and since there is no current flowing through C_{2s}, wire 2 will rise to the same potential as the shield. If the shield is grounded, then $V_2 = V_s = 0$. This provides electrostatic shielding. A similar calculation can be done for inductive coupling where the capacitive components shown in Fig. 10.13 are now replaced by inductive components. For the case shown in Fig. 10.13 the voltage induced on conductor 2 due a current flowing in conductor 1 is,

$$V_2 = M_{12} \frac{dI_1}{dt} \tag{10.27}$$

where M_{12} is the mutual inductance between the two conductors. If a shield is added as shown in Fig. 10.13, which is floating or connected to ground at only one point, then the induced voltage remains unchanged as given by Eq. (10.27). Only if the shield is connected to ground at both ends will there be a reduction of the induced voltage. Details may be found in Refs. 2 and 29.

The situation becomes considerably more complex as the frequency increases and the length of the conductor becomes comparable to the wavelength. The configuration is shown in Fig. 10.14 with terminations added. The problem is posed as follows: one conductor (the "generator wire" G) is driven by a source. What will be the induced

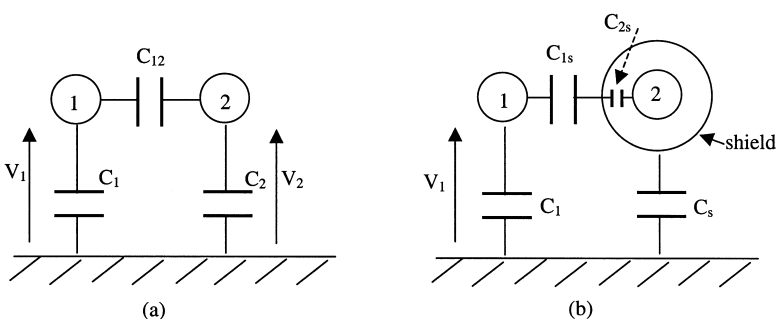


Figure 10.13 Simple electrostatic coupling models between (a) two wires and (b) two wires with shield.

© 2006 by Taylor & Francis Group, LLC

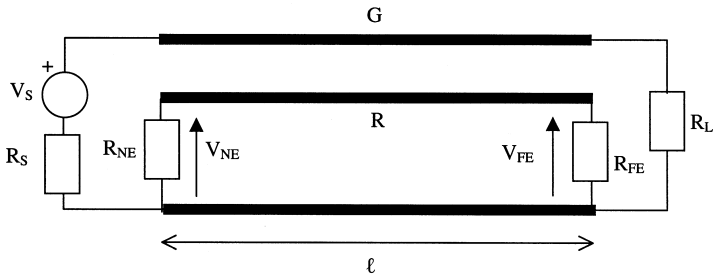


Figure 10.14 Three-wire model for studying cross talk.

voltage on the "receiver wire" R at the end near the source [near end (NE)] and far from the source [far end (FE)]? A full treatment of this problem is given in Refs. 30 and 58. We summarize here the main conclusions. For this configuration the per unit length parameter matrices are

$$[L] = \begin{bmatrix} L_G & M \\ M & L_R \end{bmatrix}$$
$$[C] = \begin{bmatrix} c_G + c_M & -c_M \\ -c_M & c_R + c_M \end{bmatrix} = \begin{bmatrix} C_G & -C_M \\ -C_M & C_R \end{bmatrix} \tag{10.28}$$

The near-end and far-end cross-talk voltages are

$$V_{\mathrm{NE}} = \frac{S}{\mathrm{Den}}\left[\frac{R_{\mathrm{NE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M\ell\left(C + \frac{j2\pi\ell/\lambda}{\sqrt{1-k^2}}\alpha_{L_G} S\right) I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell\left(C + \frac{j2\pi\ell/\lambda}{\sqrt{1-k^2}}\frac{1}{\alpha_{L_G}} S\right) V_{G_{\mathrm{dc}}}\right] \tag{10.29}$$

$$V_{\mathrm{FE}} = \frac{S}{\mathrm{Den}}\left(-\frac{R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M\ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}}\right) \tag{10.30}$$

where,

$$\mathrm{Den} = C^2 - S^2\omega^2\tau_G\tau_R\left[1 - k^2\frac{(1-\alpha_{S_G}\alpha_{L_R})(1-\alpha_{L_G}\alpha_{S_R})}{(1+\alpha_{S_R}\alpha_{L_R})(1+\alpha_{S_G}\alpha_{L_G})}\right] + j\omega CS(\tau_G + \tau_R)$$
$$C = \cos\beta\ell$$
$$S = \frac{\sin\beta\ell}{\beta\ell}$$
$$k = \frac{M}{\sqrt{L_G L_R}} = \frac{C_M}{\sqrt{C_G C_R}} \leq 1$$

© 2006 by Taylor & Francis Group, LLC

$$\tau_G = \frac{L_G \ell}{R_S + R_L} + C_G \ell \frac{R_S R_L}{R_S + R_L}$$

$$\tau_R = \frac{L_R \ell}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} + C_R \ell \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}}$$

$$V_{G_{\mathrm{dc}}} = \frac{R_L}{R_S + R_L} V_S$$

$$I_{G_{\mathrm{dc}}} = \frac{V_S}{R_S + R_L}$$

$$Z_{C_G} = \sqrt{\frac{L_G}{C_G}} \qquad Z_{C_R} = \sqrt{\frac{L_R}{C_R}}$$

$$\alpha_{S_G} = \frac{R_S}{Z_{C_G}} \qquad \alpha_{L_G} = \frac{R_L}{Z_{C_G}} \qquad \alpha_{S_R} = \frac{R_{\mathrm{NE}}}{Z_{C_R}} \qquad \alpha_{L_R} = \frac{R_{\mathrm{FE}}}{Z_{C_R}}$$

These equations are exact within the limitations of transmission line theory, i.e., TEM approximation is valid and that radiation from the line is negligible. If we assume that the lines are electrically short ($\ell \ll \lambda$) and that they are weakly coupled ($k \ll 1$), then the equations simplify to

$$\begin{aligned} V_{\mathrm{NE}} &= \frac{1}{\mathrm{Den}} \left[\frac{R_{\mathrm{NE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M \ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}} \right] \\ V_{\mathrm{NE}} &= \frac{1}{\mathrm{Den}} \left[-\frac{R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M \ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}} \right] \\ \mathrm{Den} &\simeq (1 + j\omega\tau_G)(1 + j\omega\tau_R) \end{aligned} \tag{10.31}$$

For small frequencies ($\omega\tau \ll 1$), Den $\to 1$ and the expressions above simplify further to

$$\begin{aligned} V_{\mathrm{NE}} &= \frac{R_{\mathrm{NE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M \ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}} \\ V_{\mathrm{FE}} &= -\frac{R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M \ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}} \end{aligned} \tag{10.32}$$

Both sets of simplified Eqs. (10.31) and (10.32) consist of two terms. The first term indicates inductive coupling and the second capacitive coupling. A study of these expressions permits the following general conclusions to be drawn:

Inductive coupling dominates for low-impedance loads.
Capacitive coupling dominates for high-impedance loads.
Coupling is proportional to frequency; hence, the faster the rate of change of the driving source, the higher the cross-talk levels.
Capacitive components at the near and far end are of the same magnitude and sign.
Inductive components are unequal and of opposite sign; hence, it is possible to choose the terminations to eliminate far-end cross talk.

© 2006 by Taylor & Francis Group, LLC

Cross talk at very high frequencies, where the TEM approximation is not valid and where there is substantial radiation from the line, can only be studied by numerical techniques. Another aspect of propagation on multiconductor lines which affects EMC and SI is the presence of more than one mode of propagation. In general, in a system of n lines there are $n-1$ modes of propagation. These modes may travel at different velocities and recombine at loads and discontinuities to produce distorted signals that contribute to noise. An introduction to modal propagation is given in Ref. 2 and a more complete treatment in [59,60]. Formulas for the calculation of parameters of some typical lines are given in Table 10.2.

Far-field radiative coupling refers to the coupling of external EM radiation onto circuits and the reverse effect of emission of EM radiation from circuits.

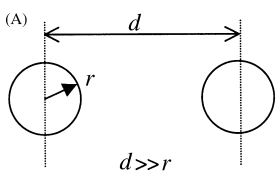
A typical problem is the calculation of voltages induced on interconnects subject to incident plane waves. A typical configuration is shown in Fig. 10.15a. There are three equivalent formulations to this problem [61]. According to the approach described in Ref. 62, the coupling to the field is described by two equivalent sources representing

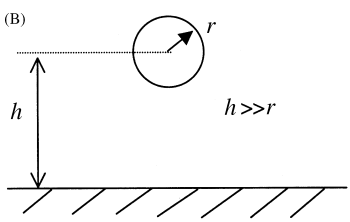
Table 10.2 Electrical Parameters of Some Common Configurations

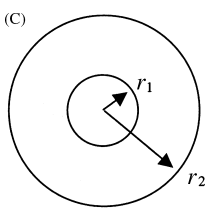
Two parallel wires (A)	$C = \frac{\pi\varepsilon}{\ln(d/r)}$ F/m	$L = \frac{\mu}{\pi}\ln\frac{d}{r}$ H/m
Wire above ground (B)	$C = \frac{2\pi\varepsilon}{\ln(2h/r)}$ F/m	$L = \frac{\mu}{2\pi}\ln\frac{2h}{r}$ H/m
Coaxial cable (C)	$C = \frac{2\pi\varepsilon}{\ln(r_2/r_1)}$ F/m	$L = \frac{\mu}{2\pi}\ln\frac{r_2}{r_1}$ H/m
Two wires above ground (D)	$[C] = \begin{bmatrix} c_{11}+c_{12} & -c_{12} \\ -c_{21} & c_{22}+c_{21} \end{bmatrix}$	$[L] = \frac{\mu}{2\pi}\begin{bmatrix} \ln\frac{2h}{r} & \ln\frac{D}{d} \\ \ln\frac{D}{d} & \ln\frac{2h}{r} \end{bmatrix}$

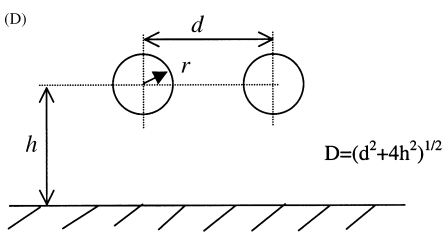
$$c_{11} = c_{22} = A\ln\frac{2h}{r} \qquad c_{12} = c_{21} = A\ln\frac{D}{d}$$

$$A = \frac{2\pi\varepsilon}{[\ln(2h/r)]^2 - [\ln(D/d)]^2}$$









© 2006 by Taylor & Francis Group, LLC

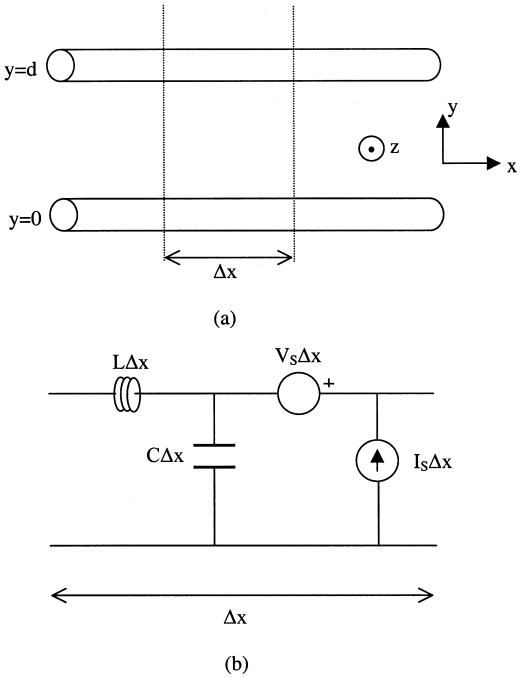


Figure 10.15 Coupling of an external field onto a two-wire line (a) and circuit model (b).

incident electric and magnetic field. The relevant modified transmission-line equations [63] are shown below:

$$\begin{aligned}\frac{dV(x)}{dx} &= -j\omega LI(x) + V_s(x)\\ \frac{dI(x)}{dx} &= -j\omega CV(x) + I_s(x)\end{aligned} \tag{10.33}$$

where L and C are the per unit length inductance and capacitance of the line and V_s and I_s are equivalent sources given by

$$\begin{aligned}V_s(x) &= j\omega\mu\int_0^d H_z^i(x,y)\,dy\\ I_s(x) &= -j\omega C\int_0^d E_y^i(x,y)\,dy\end{aligned} \tag{10.34}$$

H^i and E^i represent the incident field components. A complete treatment for different types of incident field may be found in Refs. 61–63. As an illustration we show the induced current at the two terminations of a line subject to end-fire excitation as shown in Fig. 10.16.

$$\begin{aligned}I_{\mathrm{NE}} &= j\frac{dE_0}{D}\sin\beta\ell\left(1+\frac{Z_L}{Z_C}\right)\\ I_{\mathrm{FE}} &= \frac{dE_0}{2D}\left(1-\frac{Z_S}{Z_C}\right)(1-\cos 2\beta\ell + j\sin 2\beta\ell)\end{aligned} \tag{10.35}$$

© 2006 by Taylor & Francis Group, LLC

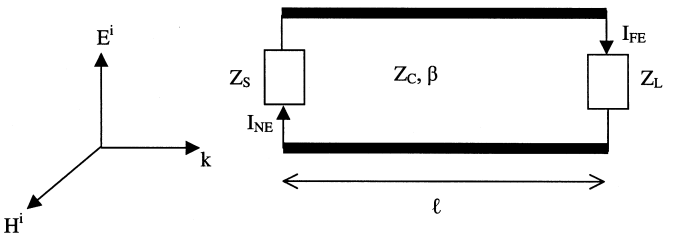


Figure 10.16 Configuration for the study of end-fire coupling.

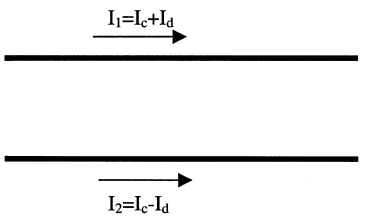


Figure 10.17 Common- and differential-mode currents on a two-wire line.

where, E_0 is the magnitude of the incident electric field and

$$D = \cos\ \beta\ell\,(Z_S + Z_L) + j \sin\ \beta\ell \left(Z_C + \frac{Z_S Z_L}{Z_C}\right)$$

Similar results for other types of excitation may be found in the references given. The reverse problem, namely, the emission of radiation from interconnects, is also important. Analytical techniques rely on a calculation of the currents that flow in an interconnect and then using this information together with antenna theory to obtain the radiated field. Of crucial importance in such calculations is the correct estimation of the current flowing in interconnects. In complex practical configurations, this current is not normally simply that calculated by transmission-line theory (differential current I_d). There is in addition a current component, which is due to a variety of mechanisms such as stray currents to nearby structures, which is described as common mode current I_c. The total current is the superposition of these two currents. The situation is shown schematically in Fig. 10.17. Only if $I_c = 0$ is the total current $I_1 = -I_2$. As the differential current components on the two wires are equal and opposite any radiation from them decays quickly with distance. In contrast, common mode current are in the same direction and make additive contributions to the radiated field. Although in general $I_c \ll I_d$ the contribution of the common mode current to radiation can dominate. Useful formulas for estimating the maximum electric field at a distance d from two parallel wires (separation s, length ℓ) are given below [30]:

$$\begin{aligned} E_{D,\max} &= 1.316 \times 10^{-14} \frac{|I_d| f^2 \ell s}{d} \\ E_{C,\max} &= 1.257 \times 10^{-6} \frac{|I_c| f \ell}{d} \end{aligned} \tag{10.36}$$

where f is the frequency.

© 2006 by Taylor & Francis Group, LLC

At high frequencies, numerical solutions are generally necessary. In some cases analytical solutions may be obtained [31].

10.4. PRACTICAL TECHNIQUES FOR THE CONTROL OF INTERFERENCE

Practical design to achieve EMC involves a series of measures to reduce emissions at source, a reduction in the efficiency of coupling paths, and improvements in immunity. Many of these techniques have been mentioned in the previous section.

As already pointed out, fast rise and fall times introduce a very wide spectrum of frequencies. Hence the slowest logic family should be used compatible with operational needs; otherwise, it is difficult to achieve EMC. Similarly, a system should be designed with the narrowest bandwidth to minimize the risk of becoming victim to noise. Many advanced systems use spread-spectrum techniques to address EMC and security problems [64]. Proper software design can also contribute to the immunity of systems to interference by including error checking and correction routines. Shielding and grounding of systems must be well thought out early in the design stage and closely monitored throughout the lifetime of a product. Whenever possible balanced cables should be used as this minimizes the flow of common-mode currents and the attendant problems of high emissions. Matching of interconnects to minimize reflections should be used whenever possible. Particular attention should be paid to the choice and installation of connectors as they tend to be the weak link in an otherwise well designed system. In critical systems, isolation techniques could be used (isolation transformers, optical links) to break interference paths. Filtering of cables on entry and exit from equipment must be considered. In particularly severe environments nonlinear limiting devices should be used to absorb high-energy pulses prior to further attenuation by filters, etc.

Positioning of circuits to minimize interference should be done carefully and segregation techniques to keep apart systems likely to interfere with each other should be considered early in the design phase. Whatever measures are taken during design to ensure EMC, they must be monitored throughout the lifetime of the product. This so-called *management of EMC* is problematic, yet it is an important aspect to consider during planning and design.

Several texts offer more detailed treatment of EMC design at the system and board level and should be consulted by designers [2,65,66].

10.5. EMC STANDARDS AND TECHNIQUES

EMC is the subject of an extensive set of international standards and national legislation to ensure that all products conform to a set of norms. Particularly important in this regard was the European EMC Directive in 1989 which gave an impetus to EMC design and certification. One can distinguish civilian standards, military standards, and company standards. Civilian standards cover a vast range of products and originate from International bodies (IEC, ITU etc.), large economic blocks, e.g., FCC from the United States and CEN from the European Community, etc. There has been a convergence of limits and procedures set by the different standards organizations, but small differences still persist in some areas. Military standards are set by the United States (MIL-STD-461D) [67], the United Kingdom (DEF-STAN 59-41) [68] and other countries. Large companies

© 2006 by Taylor & Francis Group, LLC

often impose a set of standards for internal use and for dealing with suppliers, which are designed to ensure that any national and international standards are comfortably met. There is a large range of standards available. Some standards are described as generic, i.e., they address general principles and set general limits. An example is the EN 50081 EMC Generic Emission Standards and the EN 50082, which addresses immunity. In addition, there are many product specific standards which apply to specific classes of equipment, e.g., IT equipment. In Table 10.3 we give contact addresses for the main standards bodies from which the interested reader can get up to date information.

A specialist area, which is beyond the scope of this chapter, is the setting of safe limits for human exposure to EM fields. This is important both for exposure in industrial environments and for exposure by the general public due to radio transmitters, mobile phones etc. The interested reader should consult specialist sources in this area [19,69,70].

Standards specify a test procedure and limits for emission and immunity for each class of equipment. Tests may be performed in different environments (open-area test site [71], screened or anechoic room [72], mode-stirred chamber [73], GTEM cell [74]) depending on the standard chosen. A typical test arrangement is shown in Fig. 10.18. The EUT is placed at the specified distance D from the receiving antenna Rx. The EUT is oriented for maximum received signal. Rx may be height scanned to obtain the

Table 10.3 Contact Points for EMC Standards and Codes of Practice.

International Electrotechnical Commission (IEC), 3, rue de Varembe, PO Box 131, CH-1211, Geneva, Switzerland, pubinfor@iec.ch
International Organization for Standardization (ISO). As above, central@iso.ch
European Committee for Standardization (CEN), 36, rue de Stassart, B-1050 Brussels, Belgium, infodesk@cenorm.be
CENELEC: European Committee for Electrotechnical Standardization, 35, rue de Stassart, B-1050 Brussels, Belgium, general@cenelec.be
European Telecommunications Standards Institute (ETSI), F-06921 Sophia Antipolis Cedex, France, Infocentre@etsi.fr
British Standards Institution (BSI), 389 Chiswick High Road, London, W4 4AL, UK, info@bsi-global.com
CIGRE: International Council for Large Electric Systems, 21, rue d'Artois, 75008 Paris, France, www.cigre.org
Institute of Electrical and Electronic Engineers (IEEE), 3 Park Avenue, New York 10016, USA, http://standards.ieee.org

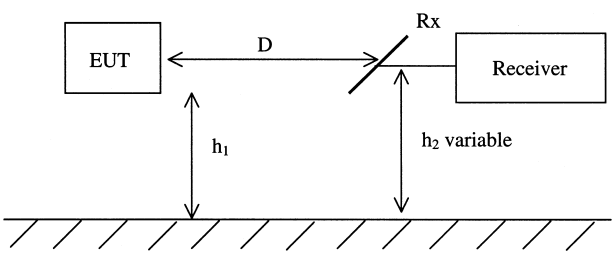


Figure 10.18 A typical arrangement for EMC testing.

© 2006 by Taylor & Francis Group, LLC

maximum signal. The electric field is calculated from the measured voltage and the antenna factor AF

$$E\text{dB}\,\mu\text{V/m} = V_{\text{rec}}\,\text{dB}\mu\text{V} + \text{AF}\,\text{dB} \tag{10.37}$$

The receiver characteristics are specified in standards (peak or quasipeak detectors [27,28]). The measured electric field over the entire frequency range is then compared with the limits specified in standards. As an example, the emission limit according to CISPR 22 measured at an open-area test site and at a distance of 30 m from the EUT is 30 dBμV/m (30–230 MHz) and 37 dBμV/m (245 MHz–1 GHz) for class A equipment (equipment for use in industrial, commercial, and business premises). The same limits apply for equipment in residential use (class B) but measured at a distance of 10 m.

Most test environments are imperfect in some way and care must be taken when taking and interpreting measurements. In screened rooms, one is confronted by the presence of room resonances. For a room of dimensions a, b, c resonances occur at,

$$f\,\text{MHz} = 150\sqrt{\left(\frac{m}{a}\right)^2+\left(\frac{n}{b}\right)^2+\left(\frac{p}{c}\right)^2} \tag{10.38}$$

where m, n, and p are integers (no more than one can be zero). Resonances make measurements difficult to interpret. Radiation damping material can be added to remove or damp resonances. This is very difficult especially at low frequencies.

In open are test sites, of major importance is the problem of ground reflections. This is illustrated in Fig. 10.19a. The signal from a transmitter Tx gets to the receiver through the direct path and a reflection from the ground. The two signal paths can be explained in terms of image theory as shown for vertical and horizontal polarizations in Fig. 10.19b,c respectively. The electrical path for the direct and image rays is different. The total signal at Rx is the superposition of these two rays and thus may vary substantially. This explains the requirement in standards that a height scan is done during tests to find the maximum field at Rx.

GTEM cells cannot be made large enough to handle very large equipment and the EUT must be small enough to avoid affecting too much the field profile inside the cell.

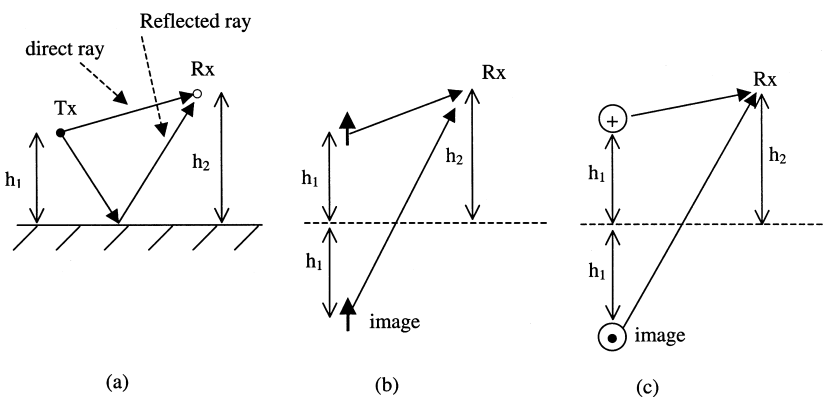


Figure 10.19 (a) Direct and reflected rays from Tx to Rx and images for (b) vertical and (c) horizontal polarization.

© 2006 by Taylor & Francis Group, LLC

The above descriptions refer to radiated emissions above 30 MHz. Below this frequency conducted emission tests are specified which consist of measuring the noise voltage across a specified mains impedance. This is provided by a line impedance stabilizing network (LISN) placed at the connection point of the EUT to the power supply. As an illustration, the CISPR 22 standard for conducted emission specifies 73 dBμV (0.5–30 MHz) for class A equipment.

10.6. OUTSTANDING PROBLEMS AND FUTURE TRENDS

There are several areas where improvements are required and further developments should be expected. In the area of testing there is a need to improve methods and procedures so that tests made in a particular environment can be referred to another environment. This will reduce the number of tests required and improve repeatability of measurements.

EMC tests at frequencies above 1 GHz are tedious and prolonged as it is difficult to scan accurately such a vast range. The directivity of emissions from equipment tends to be higher at higher frequencies. Different standards and approaches to testing may thus be required to rationalize procedures and do more meaningful tests. This is becoming increasingly important as high-speed electronics are becoming more prevalent. Along the same lines, as clock frequencies get higher EMC and SI issue become more closely related and therefore design and predictive tools should be integrated so that these two aspects may be treated concurrently throughout the design phase.

Further major developments are required to improve predictive capabilities and thus to allow iterative EMC design to take place at all stages of design. The tools required will be computer based and will incorporate sophisticated models to handle efficiently the complexities of modern systems.

Another aspect of analysis and synthesis methods in EMC is the way in which uncertainty (manufacturing and component tolerances) is dealt with. Standards, test procedures, and CAD tools are essentially based on deterministic models of systems. With increasing complexity and higher frequencies, consideration should be given to statistical techniques in the EMC characterization and design of systems [75].

There is considerable pressure on the electromagnetic spectrum to accommodate new services such as Bluetooth and TETRA [76]. The impact of these trends on interference and on general background noise levels should be considered so that the EM spectrum remains a well-managed resource for the benefit of all.

REFERENCES

1. International Electrotechnical Commission (IEC). Email: pubinfor@iec.ch
2. Christopoulos, C. *Principles and Techniques of Electromagnetic Compatibility*; CRC Press: Boca Raton, 1995.
3. Young, B. *Digital Signal Integrity*; Prentice Hall PTR: Upper Saddle River, NJ, 2001.
4. Walker, C.S. *Capacitance, Inductance, and Cross-talk Analysis*; Artech House: Boston, Norwood, MA, 1990.
5. Volakis, J.L.; Chaterjee, A.; Kempel, L.C. *Finite Element Method in Electromagnetics*; IEEE Press: New York, 1998.
6. Harrington, R.F. *Field Computation by Moment Methods*; Macmillan: New York, 1968.
7. Taflove, A. *Computational Electrodynamics: The Finite-Difference Time-Domain Method*; Artech House: Norwood, MA, 1995.

© 2006 by Taylor & Francis Group, LLC

8. Christopoulos, C. *The Transmission-Line Modeling Method: TLM*; IEEE Press: New York, 1995.
9. Balanis, C. *Antenna Theory: Analysis and Design*, 2nd Ed.; Wiley Interscience: New York, 1996.
10. Papoulis, A.; Pillai, S.U. *Probability, Random Variables and Stochastic Processes*, 4th Ed.; McGraw Hill: New York, 2002.
11. Papoulis, A. *Signal Analysis*; McGraw Hill: New York, 1977.
12. Uman, M.A.; Krider, E.P. A review of natural lightning: experimental data and modeling. IEEE Trans. EMC **1982**, *24*, 79–112.
13. Gardner, R.L.; Baker, L.; Baum, C.E.; Andersh, D.J. Comparison of lightning with public domain HEMP waveforms on the surface of an aircraft. Proc. 6th Zurich Symp. on EMC, Zurich, 1985; pp. 175–180.
14. ITU Report 670-1. Worldwide minimum external noise levels, 0.1 Hz to 100 GHz, Dusseldorf, 1990, Annex to Vol. 1.
15. Davenport, E.M.; Frank, P.J.; Thomson, J.M. Prediction of field strengths near HF transmitters. Radio Electron. Eng. **1983**, *53*, 75–80.
16. Independent Expert Group on Mobile Phones (IEGMP). Mobile Phones and Health, http://www.iegmp.org.uk, 2000.
17. Sugiura, A.; Okamura, M. Evaluation of interference generated by microwave ovens. Proc. 7th Zurich Symp. EMC, 1987; pp. 267–269.
18. Koga, R.; Wade, O.; Hiraoka, T.; Sano, H. Estimation of electromagnetic impulse noise radiated from a digital circuit board. Proc. Int. Conf. EMC, Nagoya, 1989, pp. 389–393.
19. Ran, L.; Gokani, S.; Clare, J.C.; Bradley, K.J.; Christopoulos, C. Conducted electromagnetic emissions in induction motor drives—Parts I and II. IEEE Trans. Power Electronics **1998**, *4*, 757–776.
20. Working Group 36.04. Guide on EMC in power Plants and Substations, Paris: CIGRE, 1997
21. Ma, M.T. How high is the level of EM fields radiated by an ESD. Proc. 8th Int. Zurich EMC Symp., Zurich, 1989, pp. 361–365.
22. Longmire, C.L. On the electromagnetic pulse produced by nuclear explosions. IEEE Trans. EMC **1978**, *29*, 3–13.
23. Gardner, R.L.; Baker, L.; Baum, C.E.; Andersh, D.J. Comparison of lightning with public domain HEMP waveforms on the surface of an aircraft. Proc. 6th Int. Zurich EMC Symp., Zurich, 1985, pp. 175–180.
24. ITU Report 258-5. Man-made Radio Noise. Dusseldorf, 1963-1990, Annex to Vol. VI, 1990.
25. IEEE recommended Practice for an Electromagnetic Site Survey (10 kHz–10 GHz), IEEE Std. 473, 1985.
26. Suprynowitz, V.A. *Electrical and Electronics Fundamentals*; West Publ. Company, St Paul (MN), 1987.
27. Geselowitz, D.B. Response of ideal radio noise meter to continuous sine wave, recurrent impulses, and random noise. IRE Trans. Radio Interference **1961**, 2–11.
28. CISPR Publ. 16. Specification for radio interference measuring apparatus and measurement methods.
29. Ott, H.W. *Noise Reduction Techniques in Electronic Systems*, 2nd Ed.; Wiley Interscience: New York, 1988.
30. Paul, C.R. *Introduction to Electromagnetic Compatibility*; Wiley Interscience: New York, 1992.
31. Tesche, F.M.; Ianoz, M.V.; Karlsson, T. *EMC Analysis Methods and Computational Models*; Wiley Interscience: New York, 1997.
32. Cooley, W.W. Low-frequency shielding effectiveness of nonuniform enclosures. IEEE Trans. EMC **1968**, *10*, 34–43.
33. King, L.V. Electromagnetic shielding at radio frequencies. Philos. Mag. **1993**, *15*, 201–223.
34. Thomas, A.K. Magnetic shielding enclosure design in the dc and VLF region. IEEE Trans. EMC **1968**, *10*, 142–152.
35. Lee, K.S.H. Electromagnetic shieding. In *Recent Advances in Electromagnetic Theory*; Kritikos, H.N., Jaggard, D.L., Eds.; Springer-Verlag: New York, 1990.

© 2006 by Taylor & Francis Group, LLC

36. Schelkunoff, S.A. *Electromagnetic Waves*; Van Nostrand: Toronto, 1943.
37. Kaden, H. *Wirbelstrome und Schirmung in der Nachrichtentechnik*, 2nd Ed.; Springer-Verlag: New York, 1959.
38. Trenkic, V.; Duffy, A.P.; Benson, T.M.; Christopoulos, C. Numerical simulation of penetration and coupling using the TLM method. Proc. EMC Symp., Rome, 1994, pp. 321–326.
39. Trenkic, V.; Christopoulos, C.; Benson, T.M. Numerical simulation of polymers and other materials for electronic shielding applications. Proc. Polymat 94, London, 1994, 384–387.
40. Bethe, H.A. Theory of diffraction by small holes. Phys. Rev. **1944**, *66*, 163–182.
41. Cohn, S.B. Electric polarizability of apertures of arbitrary shape. Proc. IRE, 1952, pp. 1069–1071.
42. Robinson, M.P.; Turner, J.D.; Thomas, D.W.P.; Dawson, J.F.; Ganley, M.D.; Marvin, A.C.; Porter, S.J.; Benson, T.M.; Christopoulos, C. Shielding effectiveness of a rectangular enclosure with a rectangular aperture. Electronics Lett. **1996**, *32*, 1559–1560.
43. Sewell, P.; Turner, J.D.; Robinson, M.P.; Thomas, D.W.P.; Benson, T.M.; Christopoulos, C.; Dawson, J.F.; Ganley, M.D.; Marvin, A.C.; Porter, S.J. Comparison of analytic numerical and approximate models for shielding effectiveness with measurements. IEE Proc.-Sci. Meas. Technol. **1998**, *145*, 61–66.
44. Thomas, D.W.P.; Denton, A.; Konefal, T.; Benson, T.M.; Christopoulos, C.; Dawson, J.F.; Marvin, A.C.; Porter, S.J. Characterization of the shielding effectiveness of loaded equipment cabinets. EMC York 99, IEE Conf. Publ. 464, York, 1999, pp. 89–94.
45. Thomas, D.W.P.; Denton, A.; Konefal, T.; Benson, T.M.; Christopoulos, C.; Dawson, J.F.; Marvin, A.C.; Porter, S.J.; Sewell, P. Model of the EM fields inside a cuboidal enclosure populated by conducting planes or printed-circuit boards. IEEE Trans. EMC **2001**, *43*, 161–169.
46. Gupta, K.C.; Garg, R.; Bahl, I.J. *Microstrip Lines and Slotlines*; Artech House: Norwood, MA, 1979, Chap. 7.
47. De Smedt, R.; De Moerloose, J.; Criel, S.; De Zutter, D.; Olyslager, F.; Laermans, E.; Wallyn, W.; Lietaert, N. Approximate simulation of the shielding effectiveness of a rectangular enclosure with a grid wall. Proc. IEEE Int. Conf. EMC, Denver, 1998, pp. 1030–1034.
48. Kraft, C.H. Modeling leakage through finite apertures with TLM. Proc. IEEE Int. Symp. EMC, Chicago, 1994, pp. 73–76.
49. Podlozny, V.; Paul, J.; Christopoulos, C. Efficient calculation of the shielding effectiveness of equipment cabinets in full-field numerical models. Proc. EMC Europe 2002, Sorrento, 2002, pp. 853–857.
50. Podlozny, V.; Christopoulos, C.; Paul, J. Efficient description of fine features using digital filters in time-domain computational electromagnetics. IEE Proc. Sci. Meas. Technol. **2002**, *149*, 254–257.
51. Tang, T.G.; Tieng, Q.M.; Gunn, M.W. Equivalent circuit of a dipole antenna using frequency-independent lumped elements. IEEE Trans. AP **1993**, *41*, 100–103.
52. Thomas, D.W.P.; Denton, A.; Benson, T.M.; Christopoulos, C.; Paul, J.; Konefal, T.; Dawson, J.F.; Marvin, A.C.; Porter, S.J. Electromagnetic coupling to an enclosure via a wire penetration. Proc. IEEE Int. Symp. EMC, Montreal, 2001, pp. 183–188.
53. Sarto, M.S.; Scarlatti, A. Combined FDTD-TL modelling of a transmission line crossing a metallic box. Proc. 4th European Symp. EMC, Brugge, 2000, pp. 239–244.
54. Vance, E.F. Shielding effectiveness of braided-wire shields. IEEE Trans. EMC **1975**, *17*, 71–77.
55. Casey, K.F. EMP coupling through cable shields. IEEE Trans. EMC **1978**, *20*, 100–106.
56. Hoeft, L.O.; Hofstra, J.S.; Peel, R.J. Experimental evidence for purpoising coupling and optimization in braided cables. Proc. 8th Zurich Int. EMC Symp., Zurich, 1989, pp. 505–509.
57. Benson, F.A.; Cudd, P.A.; Tealby, J.M. Leakage from coaxial cables. IEE Proc. **1992**, *A 139*, 285–302.
58. Paul, C.R. Solution of the transmission-line equations for three conductor lines in homogeneous media. IEEE Trans. EMC **1978**, *20*, 216–222.
59. Djordjevic, A.R.; Sarkar, T.K.; Harrington, R.F. Time-domain response of multiconductor transmission lines. Proc. IEEE **1987**, *75*, 743–764.

© 2006 by Taylor & Francis Group, LLC

60. Paul, C.R. Frequency response of multiconductor transmission lines illuminated by an electromagnetic field. IEEE Trans. EMC **1976**, *18*, 183–190.
61. Nucci, C.A.; Rachidi, F. On the contribution of the electromagnetic field components in field-to-transmission line interaction. IEEE Trans. EMC **1995**, *37*, 505–508.
62. Abraham, R.T.; Paul, C.R. Basic EMC technology advancement for C^3 systems—coupling of EM fields onto transmission lines. RADC-TR-82-286, Vol. IVA, 1982.
63. Paul, C.R. *Analysis of Multiconductor Transmission Lines*; Wiley Interscience: New York, 1994.
64. Dixon, R.C. *Spread Spectrum Systems*; Wiley Interscience: New York, 1984.
65. Weston, D.A. *Electromagnetic Compatibility—Principles and Applications*; Marcel Dekker: New York, 1991.
66. Williams, T. *EMC for Systems and Installations*; Oxford: Newnes, 1999.
67. Department of Defence, Washington, DC. Requirement for the Control of Electromagnetic Interference Emissions and Susceptibility. MIL-STD-461D, 1993.
68. Ministry of Defence, Glasgow, UK. Electromagnetic Compatibility. DEF-STAN 59-41, 1988.
69. International Commission on Nonionizing Radiation Protection (ICNIRP). Guidelines on limits of exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). Health Phys **1998**, *74*, 494–522.
70. Repacholi, M.H. Assessment of the health effects of EMF exposure. Radio Sci. Bull. **2002**, *301*, 14–24.
71. Smith, A.A. Standard-site method for determining antenna factors. IEEE Trans. EMC **1982**, *24*, 316–322.
72. Christopoulos, C.; Paul, J.; Thomas, D.W.P. Absorbing materials and damping of screened rooms for EMC testing. Proc. Int. Symp. EMC, Tokyo, 1999, pp. 504–507.
73. Ma, M.T. Understanding reverberating chambers as an alternative facility for EMC testing. J. Electromagnetic Waves Appl. **1988**, *2*, 339–351.
74. Garbe, H.; Hansen, D. The GTEM cell concept: application of this new test environment to emission and susceptibility measurements. Proc. 7th Int. Conf. EMC, York, 1990, pp. 152–156.
75. Holland, R.; St John, R. *Statistical Electromagnetics*; Taylor and Francis: Philadelphia, PA, 1999.
76. Wesolowski, K. *Mobile Communication Systems*; Wiley Interscience: New York, 2002.

© 2006 by Taylor & Francis Group, LLC

Appendix A
Some Useful Constants

Permittivity of free space (ε_0) $= 8.854 \times 10^{-12}$ F/m
Permeability of free space (μ_0) $= 4\pi \times 10^{-7}$ H/m
Speed of electromagnetic waves in free space (c) $= 3 \times 10^{8}$ m/s
Impedance of free space (Z_0 or η_0) $= 376.7\ \Omega$
Boltzmann's constant (k) $= 1.38 \times 10^{-23}$ J/K
Charge of electron (e or q_e) $= -1.602 \times 10^{-19}$ C

© 2006 by Taylor & Francis Group, LLC

Appendix B
Some Units and Conversions

Quantity	SI[a] unit	Conversion factor
Length	meter (m)	= 39.37 in
Mass	kilogram (kg)	= 2.21 pound-mass (lb_m)
Time	second (s)	
Frequency	hertz (Hz)	= 1 cycle/s
Force	newton (N)	= 0.2248 pound-force (lb_f)
Charge	coulomb (C)	
Charge density	coulomb/meter^3 (C/m^3)	
Current	ampere (A)	
Current density	ampere/ meter^2 (A/m^2)	
Electric field	volt/meter (V/m)	
Electric flux density	coulomb/meter^2 (C/m^2)	
Magnetic field	ampere/meter (A/m)	
Magnetic flux density	tesla (T)	= 10,000 G
	or weber/meter^2 (Wb/m^2)	= 10,000 G
Resistance	ohm (Ω)	
Conductivity	siemens/meter (S/m)	
	or mho/m	
Capacitance	farad (F)	
Permittivity	farad/meter (F/m)	
Inductance	henry (H)	
Permeability	henry/meter (H/m)	

[a]SI = International System of Units.

© 2006 by Taylor & Francis Group, LLC

Appendix C
Review of Vector Analysis and Coordinate Systems

Since the formulation and application of various electromagnetic laws is greatly facilitated by the use of vector analysis, this appendix presents a concise review of vector analysis and the principal coordinate systems.

C.1. SCALARS AND VECTORS

A *scalar* quantity can be expressed as a single real number. (It can be positive, negative, or zero.) For example, voltage and current are scalar quantities. In ac analysis it is mathematically convenient to use *phasors* to represent sinusoidally varying voltages and currents. Phasors are referred to as *complex scalars*, since they require complex numbers (either magnitude and phase or real and imaginary parts) for their specification.

A *vector* quantity (e.g., the electric field) requires both a magnitude and a direction for its specification. The magnitude is always positive (it may be zero).

C.2. THE RECTANGULAR COORDINATE SYSTEM

The rectangular coordinate system (Fig. C.1a) locates a point P in three-dimensional space by assigning to it the coordinates (x_1, y_1, z_1) within a frame of reference defined by three mutually orthogonal (perpendicular) axes: the x axis, the y axis, and the z axis. It is conventional to choose a *right-handed* coordinate system (and we will do so throughout this handbook). This choice simply means that if we first point the fingers of the right hand along the x axis and then curl them to point along the y axis, the extended thumb will align with the z axis.

To deal with vectors, we define a set of three unit vectors $\mathbf{a}_x$, $\mathbf{a}_y$, and $\mathbf{a}_z$ (each with a magnitude equal to one) aligned with (parallel to) the three axes. An arbitrary vector $\mathbf{A}$ may now be expressed as $\mathbf{A} = A_x\ \mathbf{a}_x + A_y\ \mathbf{a}_y + A_z\ \mathbf{a}_z$, where A_x, A_y, A_z are said to be its scalar components along the three axes. The vector $\mathbf{A}$ has a magnitude $A = [A_x^2 + A_y^2 + A_z^2]^{1/2}$. Figure C.1b shows a differential volume $dV = dx\,dy\,dz$. The surfaces have differential areas, ds, of $dx\,dy$, $dy\,dz$, and $dz\,dx$.

C.3. SCALAR AND VECTOR FIELDS

The concepts of scalars and vectors introduced in Sec. C.1 can be extended to define scalar and vector fields. A scalar field associates a scalar quantity with every point in a

© 2006 by Taylor & Francis Group, LLC

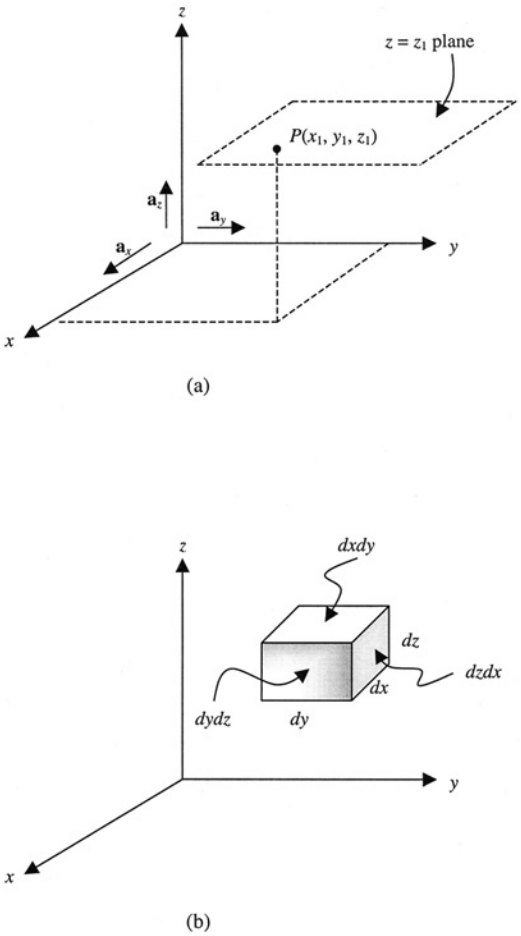


Figure C.1 The rectangular coordinate system: (a) the coordinates of a point and the unit vectors and (b) differential elements.

region of space. If we set up a rectangular coordinate system to identify various points in the 3D space in a room, we may describe the temperature distribution (scalar field) as some function $T = f(x, y, z)$ so that at the point (x_1, y_1, z_1) the temperature $T(x_1, y_1, z_1)$ is given by the value of the function $f(x_1, y_1, z_1)$. In a similar fashion, if we associate a vector with every point in a region, we will have a vector field. In the rectangular coordinate system, we can write a vector field in terms of its three components, each of which is a scalar field. For example, the velocity distribution in a river may be expressed as $\mathbf{v} = v_x(x, y, z)\mathbf{a}_x + v_y(x, y, z)\mathbf{a}_y + v_z(x, y, z)\mathbf{a}_z$.

C.4. VECTOR ADDITION AND SUBTRACTION

Two vectors **A** and **B** may be added together graphically by the familiar *parallelogram rule* shown in Fig. C.2. The addition can also be performed by adding the corresponding components of the two vectors.

© 2006 by Taylor & Francis Group, LLC

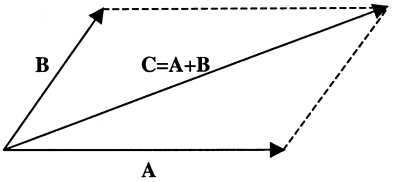


Figure C.2 Vector addition via the parallelogram rule.

If $\mathbf{A} = A_x\mathbf{a}_x + A_y\mathbf{a}_y + A_z\mathbf{a}_z$ and $\mathbf{B} = B_x\mathbf{a}_x + B_y\mathbf{a}_y + B_z\mathbf{a}_z$, their sum is a vector **C**, given as

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = (A_x + B_x)\mathbf{a}_x + (A_y + B_y)\mathbf{a}_y + (A_z + B_z)\mathbf{a}_z$$

Vector addition always obeys the following laws:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad \text{(commutative)}$$

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C} \qquad \text{(associative)}$$

Vector subtraction, $\mathbf{A} - \mathbf{B}$, is accomplished by reversing the direction of **B** to obtain another vector $-\mathbf{B}$ and then adding it to the vector **A**. Thus we have

$$\mathbf{D} = \mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B})$$

or

$$\mathbf{D} = (A_x - B_x)\mathbf{a}_x + (A_y - B_y)\mathbf{a}_y + (A_z - B_z)\mathbf{a}_z$$

where **A** and **B** have been expressed in terms of their rectangular components.

In dealing with vector fields, it is important to realize that we should be adding and subtracting only those vectors that are defined at the same point in space.

C.5. POSITION AND DISTANCE VECTORS

The position vector associated with a point P, which has the rectangular coordinates (x_1, y_1, z_1), is the vector extending from the origin $O(0, 0, 0)$ to the point P. It may be expressed as (Fig. C.3)

$$\mathbf{OP} = x_1\mathbf{a}_x + y_1\mathbf{a}_y + z_1\mathbf{a}_z \tag{C.1}$$

The distance vector **PQ** extends from the point $P(x_1, y_1, z_1)$ to the point $Q(x_2, y_2, z_2)$ and can be expressed as

$$\begin{aligned}\mathbf{PQ} &= \mathbf{OQ} - \mathbf{OP} \\ &= (x_2\mathbf{a}_x + y_2\mathbf{a}_y + z_2\mathbf{a}_z) - (x_1\mathbf{a}_x + y_1\mathbf{a}_y + z_1\mathbf{a}_z) \\ &= (x_2 - x_1)\mathbf{a}_x + (y_2 - y_1)\mathbf{a}_y + (z_2 - z_1)\mathbf{a}_z\end{aligned} \tag{C.2}$$

© 2006 by Taylor & Francis Group, LLC

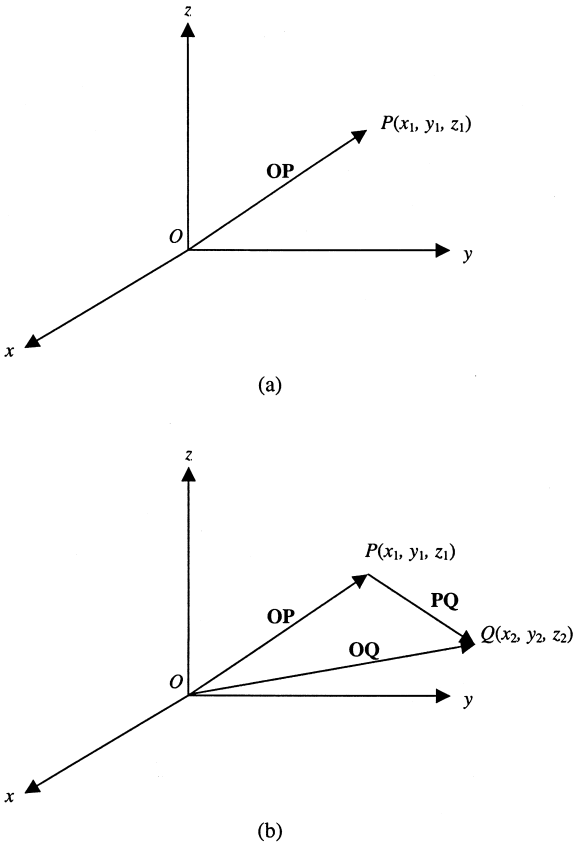


Figure C.3 (a) The position vector **OP** extends from the origin O to the point P. (b) The distance vector **PQ** extends from P to Q.

The scalar distance PQ is given by the magnitude of the vector **PQ**. Thus,

$$PQ = |\mathbf{PQ}| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{C.3}$$

C.6. VECTOR DIVISION AND MULTIPLICATION

The operation **A**/**B** is *not* defined. However, a vector can be divided by a scalar.

Two forms of vector to vector multiplication are useful in our work.

C.6.1. Scalar (Dot) Product

The scalar product of the two vectors **A** and **B** is represented symbolically as **A** • **B** (hence the alternate name, the *dot product*).

$$\mathbf{A} \bullet \mathbf{B} = |\mathbf{A}||\mathbf{B}| \cos \theta_{AB} \tag{C.4}$$

© 2006 by Taylor & Francis Group, LLC

where θ_{AB} is the smaller angle between **A** and **B**. Also,

$$\mathbf{A} \bullet \mathbf{B} = A_x B_x + A_y B_y + A_z B_z \tag{C.5}$$

The scalar product is *commutative*, i.e., $\mathbf{A} \bullet \mathbf{B} = \mathbf{B} \bullet \mathbf{A}$. Also note that

$$\mathbf{A} \bullet \mathbf{a}_x = A_x$$

$$\mathbf{A} \bullet \mathbf{a}_y = A_y$$

$$\mathbf{A} \bullet \mathbf{a}_z = A_z$$

C.6.2. Vector (Cross) Product

The vector product between **A** and **B** is a vector represented as $\mathbf{A} \times \mathbf{B}$ and is given by

$$\mathbf{A} \times \mathbf{B} = |\mathbf{A}||\mathbf{B}||\sin\theta_{AB}|\mathbf{a}_n \tag{C.6}$$

where θ_{AB} is the smaller angle between **A** and **B**, and $\mathbf{a}_n$ is a unit vector normal to the plane containing **A** and **B**. (Since each plane has two normal vectors, it is important to note that $\mathbf{a}_n$ is the one obtained by the right-hand rule. If the fingers of the right hand are extended in the direction of **A** and then curled towards vector **B**, the direction of the outstretched thumb is the direction of $\mathbf{a}_n$.)

Also,

$$\mathbf{A} \times \mathbf{B} = (A_y B_z - A_z B_y)\mathbf{a}_x + (A_z B_x - A_x B_z)\mathbf{a}_y + (A_x B_y - A_y B_x)\mathbf{a}_z \tag{C.7}$$

C.7. The Cylindrical Coordinate System

While much of our work is carried out conveniently in the familiar rectangular coordinate system (introduced in Sec. C.2), some physical situations have a natural symmetry which makes the cylindrical coordinate system easier to use. Examples include a coaxial cable and an optical fiber.

The cylindrical coordinate system we will use is a natural extension of the two dimensional (xy plane) polar coordinates (ρ,ϕ) to three dimensions (ρ,ϕ,z). Figure C.4 shows the geometric relationship between the rectangular coordinates (x, y, z) of a point P and its cylindrical coordinates (ρ,ϕ,z). You will notice that the z coordinate is common to both systems, while ρ and ϕ are related to x and y as follows

$$\rho = +\sqrt{x^2 + y^2} \qquad \phi = \tan^{-1}\left(\frac{y}{x}\right) \tag{C.8}$$

$$x = \rho\cos\phi \qquad y = \rho\sin\phi, \tag{C.9}$$

where ρ may be thought of as the "*horizontal*" radial distance from the origin to the point P, while ϕ measures the "*azimuthal*" angle from the x axis in a counterclockwise (toward the y axis) direction.

Just as, in the rectangular coordinate system, a point $P(x_1, y_1, z_1)$ corresponds to the intersection of the three mutually orthogonal *planar* surfaces: $x = x_1$, $y = y_1$, $z = z_1$,

© 2006 by Taylor & Francis Group, LLC

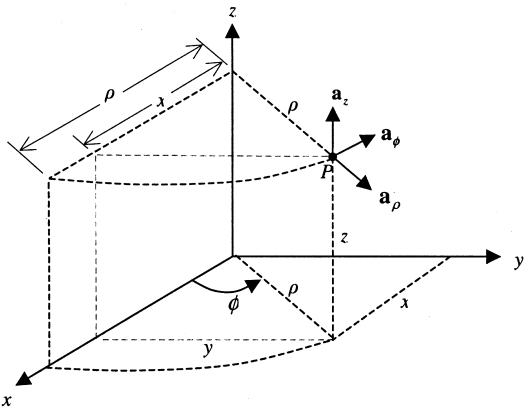


Figure C.4 The cylindrical coordinate system.

in the cylindrical system, $P(\rho_1,\phi_1,z_1)$ is located at the intersection of the three orthogonal surfaces:

$$\rho = \rho_1 \qquad \text{(cylinder)}$$
$$\phi = \phi_1 \qquad \text{(plane)}$$
$$z = z_1 \qquad \text{(plane)}$$

A vector $\mathbf{A}$ may be expressed in the cylindrical system as

$$\mathbf{A} = A_\rho \mathbf{a}_\rho + A_\phi \mathbf{a}_\phi + A_z \mathbf{a}_z \tag{C.10}$$

with $|\mathbf{A}| = (A_\rho^2 + A_\phi^2 + A_z^2)^{1/2}$, where $\mathbf{a}_\rho$, $\mathbf{a}_\phi$, and $\mathbf{a}_z$, are mutually orthogonal unit vectors as shown in Fig. C.5. $\mathbf{a}_\rho$ points in the direction of increasing "horizontal" radial distance ρ, $\mathbf{a}_\phi$ also lies in a "horizontal" plane (parallel to the xy plane) and points in the direction of increasing ϕ, and finally $\mathbf{a}_z$ is parallel to the positive z axis (as before). Also note the right-hand rule relationship among $\mathbf{a}_\rho$, $\mathbf{a}_\phi$, and $\mathbf{a}_z$, i.e.,

$$\begin{aligned} \mathbf{a}_\rho \times \mathbf{a}_\phi &= \mathbf{a}_z \\ \mathbf{a}_\phi \times \mathbf{a}_z &= \mathbf{a}_\rho \\ \mathbf{a}_z \times \mathbf{a}_\rho &= \mathbf{a}_\phi \end{aligned} \tag{C.11}$$

A differential volume element dV in the cylindrical coordinate system is

$$dV = (d\rho)(\rho d\phi)(dz) \tag{C.12}$$

Note that $\rho d\phi$ (and not $d\phi$ by itself) represents an incremental distance in the direction of $\mathbf{a}_\phi$, since $d\phi$ is a dimensionless angular measure.

C.8. THE SPHERICAL COORDINATE SYSTEM

If one wishes to analyze the scattering of microwave radar signals from raindrops or the electromagnetic interaction between a cell phone and the human head, the spherical

© 2006 by Taylor & Francis Group, LLC

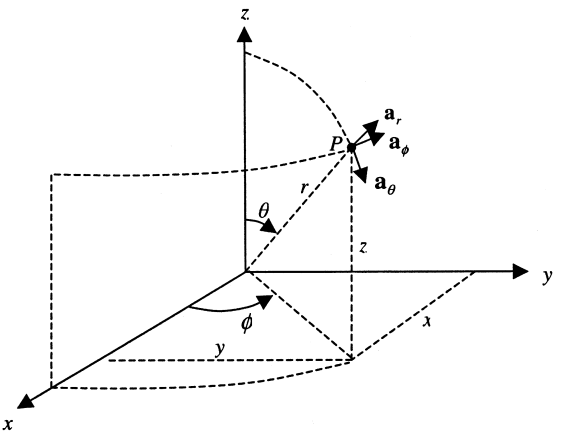


Figure C.5 The spherical coordinate system.

coordinate system may facilitate setting up the mathematical problem. In the spherical coordinate system (Fig. C.5), a point $P(r_1,\theta_1,\phi_1)$ lies at the intersection of the three mutually orthogonal surfaces

$$r = r_1 \qquad \text{(sphere)}$$
$$\theta = \theta_1 \qquad \text{(cone)}$$
$$\phi = \phi_1 \qquad \text{(plane)}$$

r represents the three-dimensional distance between the origin and the point, θ $(0 \le \theta < \pi)$ is the "*elevation*" angle measured from the positive z axis, and ϕ $(0 \le \phi < 2\pi)$ is the "*azimuthal*" angle measured from the positive x axis (as in the cylindrical coordinate system). They are related to (x, y, z) as follows:

$$r = [x^2 + y^2 + z^2]^{1/2} \qquad \theta = \cos^{-1}\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right) \qquad \phi = \tan^{-1}\left(\frac{y}{x}\right) \tag{C.13}$$

$$x = r\sin\theta\cos\phi \qquad y = r\sin\theta\sin\phi \qquad z = r\sin\theta \tag{C.14}$$

A vector $\mathbf{A}$ in the spherical coordinate system is expressed as

$$\mathbf{A} = A_r\mathbf{a}_r + A_\theta\mathbf{a}_\theta + A_\phi\mathbf{a}_\phi \tag{C.15}$$

with $|\mathbf{A}| = (A_r^2 + A_\theta^2 + A_\phi^2)^{1/2}$.

The unit vectors $\mathbf{a}_r$, $\mathbf{a}_\theta$, and $\mathbf{a}_\phi$, are mutually orthogonal and follow the right-hand rule relationship embedded in

$$\mathbf{a}_r \times \mathbf{a}_\theta = \mathbf{a}_\phi$$
$$\mathbf{a}_\theta \times \mathbf{a}_\phi = \mathbf{a}_r \tag{C.16}$$
$$\mathbf{a}_\phi \times \mathbf{a}_r = \mathbf{a}_\theta$$

© 2006 by Taylor & Francis Group, LLC

A differential volume element dV is written as

$$dV = (dr)(r\,d\theta)(r\sin\theta\,d\phi) = r^2\sin\theta\,dr\,d\theta\,d\phi \tag{C.17}$$

C.9. COORDINATE AND VECTOR TRANSFORMATION

In working with the various coordinate systems, we may need to convert parameters given in one coordinate system into parameters in another coordinate system.

Vector Interconversion Strategy

The most common conversions in practice are those between rectangular and cylindrical coordinate systems and those between rectangular and spherical coordinate systems. Both types can be accomplished easily with the help of Table C.1.

Example C.1. Vector transformation (rectangular to cylindrical): Convert $\mathbf{A} = z\mathbf{a}_x + x\mathbf{a}_y$ to cylindrical coordinates.

We start by writing $\mathbf{A} = A_\rho\mathbf{a}_\rho + A_\phi\mathbf{a}_\phi + A_z\mathbf{a}_z$. Then

$$\begin{aligned} A_\rho &= \mathbf{A}\bullet\mathbf{a}_\rho \\ &= (z\mathbf{a}_x + x\mathbf{a}_y)\bullet\mathbf{a}_\rho \\ &= z(\mathbf{a}_x\bullet\mathbf{a}_\rho) + x(\mathbf{a}_y\bullet\mathbf{a}_\rho) \\ &= z\cos\phi + (\rho\cos\phi)(\sin\phi) \\ A_\phi &= \mathbf{A}\bullet\mathbf{a}_\phi \\ &= (z\mathbf{a}_x + x\mathbf{a}_y)\bullet\mathbf{a}_\phi \\ &= z(\mathbf{a}_x\bullet\mathbf{a}_\phi) + x(\mathbf{a}_y\bullet\mathbf{a}_\phi) \\ &= z(-\sin\phi) + (\rho\cos\phi)(\cos\phi) \end{aligned}$$

Table C.1 Unit Vector Transformation

(a) Dot products of unit vectors (rectangular/cylindrical)

	$\mathbf{a}_\rho$	$\mathbf{a}_\phi$	$\mathbf{a}_z$
$\mathbf{a}_x$	$\cos\phi$	$-\sin\phi$	0
$\mathbf{a}_y$	$\sin\phi$	$\cos\phi$	0
$\mathbf{a}_z$	0	0	1

(b) Dot products of unit vectors (rectangular/spherical)

	$\mathbf{a}_r$	$\mathbf{a}_\theta$	$\mathbf{a}_\phi$
$\mathbf{a}_x$	$\sin\theta\cos\phi$	$\cos\theta\cos\phi$	$-\sin\phi$
$\mathbf{a}_y$	$\sin\theta\sin\phi$	$\cos\theta\sin\phi$	$\cos\phi$
$\mathbf{a}_z$	$\cos\theta$	$-\sin\theta$	0

© 2006 by Taylor & Francis Group, LLC

$$\begin{aligned} A_z &= \mathbf{A} \bullet \mathbf{a}_z \\ &= (z\mathbf{a}_x + x\mathbf{a}_y) \bullet \mathbf{a}_z \\ &= z(\mathbf{a}_x \bullet \mathbf{a}_z) + x(\mathbf{a}_y \bullet \mathbf{a}_z) \\ &= z \cdot 0 + x \cdot 0 \end{aligned}$$

Therefore,

$$\mathbf{A} = (z\cos\phi + \rho\cos\phi\sin\phi)\mathbf{a}_\rho + (-z\sin\phi + \rho\cos^2\phi)\mathbf{a}_\phi$$

Example C.2. Vector transformation (spherical to rectangular): Convert $\mathbf{E} = E_o/(r^2)\,\mathbf{a}_r$ into rectangular coordinates.

We start by writing $\mathbf{E} = E_x\mathbf{a}_x + E_y\mathbf{a}_y + E_z\mathbf{a}_z$. Then,

$$\begin{aligned} E_x = \mathbf{E} \bullet \mathbf{a}_x &= \left(\frac{E_o}{r^2}\right)(\mathbf{a}_r \bullet \mathbf{a}_x) = \frac{E_o}{r^2}\sin\theta\cos\phi = \frac{E_o}{r^2}\frac{r\sin\theta\cos\phi}{r} = \frac{E_o}{r^3}r\sin\theta\cos\phi \\ &= \frac{E_0 x}{(x^2+y^2+z^2)^{3/2}} \end{aligned}$$

Similarly,

$$E_y = \frac{E_o y}{(x^2+y^2+z^2)^{3/2}} \qquad \text{and} \qquad E_z = \frac{E_o z}{(x^2+y^2+z^2)^{3/2}}$$

C.10. VECTOR DIFFERENTIAL OPERATORS

Maxwell equations and the associated relationships are expressed in terms of vector differential operators. Therefore, we have tabulated the expressions for the various differential operators in all the coordinate systems below:

C.10.1. Divergence

Rectangular

$$\nabla \bullet \mathbf{D} = \frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z}$$

Cylindrical

$$\nabla \bullet \mathbf{D} = \frac{1}{\rho}\frac{\partial}{\partial\rho}(\rho D_\rho) + \frac{1}{\rho}\frac{\partial D_\phi}{\partial\phi} + \frac{\partial D_z}{\partial z}$$

Spherical

$$\nabla \bullet \mathbf{D} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 D_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(D_\theta\sin\theta) + \frac{1}{r\sin\theta}\frac{\partial D_\phi}{\partial\phi}$$

© 2006 by Taylor & Francis Group, LLC

C.10.2. Gradient

Rectangular

$$\nabla V = \frac{\partial V}{\partial x}\mathbf{a}_x + \frac{\partial V}{\partial y}\mathbf{a}_y + \frac{\partial V}{\partial z}\mathbf{a}_z$$

Cylindrical

$$\nabla V = \frac{\partial V}{\partial \rho}\mathbf{a}_\rho + \frac{1}{\rho}\frac{\partial V}{\partial \phi}\mathbf{a}_\phi + \frac{\partial V}{\partial z}\mathbf{a}_z$$

Spherical

$$\nabla V = \frac{\partial V}{\partial r}\mathbf{a}_r + \frac{1}{r}\frac{\partial V}{\partial \theta}\mathbf{a}_\theta + \frac{1}{r\sin\theta}\frac{\partial V}{\partial \phi}\mathbf{a}_\phi$$

C.10.3. Curl

Rectangular

$$\nabla \times \mathbf{E} = \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\mathbf{a}_x + \left(\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}\right)\mathbf{a}_y + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right)\mathbf{a}_z$$

Cylindrical

$$\nabla \times \mathbf{E} = \left(\frac{1}{\rho}\frac{\partial E_z}{\partial \phi} - \frac{\partial E_\phi}{\partial z}\right)\mathbf{a}_\rho + \left(\frac{\partial E_\rho}{\partial z} - \frac{\partial E_z}{\partial \rho}\right)\mathbf{a}_\phi + \frac{1}{\rho}\left(\frac{\partial}{\partial \rho}\rho E_\phi - \frac{\partial E_\rho}{\partial \phi}\right)\mathbf{a}_z$$

Spherical

$$\nabla \times \mathbf{E} = \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial \theta}\left(E_\phi \sin\theta\right) - \frac{\partial E_\theta}{\partial \phi}\right]\mathbf{a}_r + \frac{1}{r}\left(\frac{1}{\sin\theta}\frac{\partial E_r}{\partial \phi} - \frac{\partial}{\partial r}rE_\phi\right)\mathbf{a}_\theta + \frac{1}{r}\left(\frac{\partial}{\partial r}rE_\theta - \frac{\partial E_r}{\partial \theta}\right)\mathbf{a}_\phi$$

© 2006 by Taylor & Francis Group, LLC